ТАРТУСКИЙ
ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

# РАСПОЗНАВАНИЕ ОБРАЗОВ

( Материалы конференции )

ТАРТУ 1972

ТАРТУСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

# РАСПОЗНАВАНИЕ ОБРАЗОВ

( Материалы конференции )

ТАРТУ 1972

## ОТ РЕДАКЦИИ

В ноябре 1970 года кафедрой математической статистики и программирования и Лабораторией социологии Тартуского Государственного Университета был проведен научный семинар по распознаванию образов.

В семинаре, который проводился в историческом местечке Сангасте (на юге Эстонии), участвовало более 50 человек - математики, кибернетики, социологи и представители многих других специальностей. В течение четырех дней было заслушано около 25 докладов, посвященных самым разным вопросам методов распознавания образов и их применения в различных областях науки.

Ученый совет семинара, который был избран на месте, оценил семинар как вполне удавшийся и рекомендовал и в дальнейшем проводить подобные встречи. Ученый совет предложил организаторам семинара выпустить сборник материалов семинара, так как те идеи и решения, которые были высказаны в ходе выступлений и дискуссий, достойны внимания более широкой аудитории.

В настоящем сборнике публикуется часть докладов, прочитанных на семинаре в Сангасте. К сожалению, в сборник не попало несколько весьма интересных статей, так как ряд докладчиков не прислало своих докладов для опубликования.

Выражаем надежду, что настоящий сборник будет полезным для многочисленных исследователей, интересующихся методами распознавания образов и их применением.

# МЕТОДЫ РАСПОЗНАВАНИЯ ОБРАЗОВ В СОЦИОЛОГИИ

И.Б. Мучник

(Москва)

Данная работа посвящена обзору основных направлений применения методов распознавания образов в социологии. При этом не ставилась задача разбора уже проведенных исследований. Цель работы состояла в выделении специфических особенностей возникающих в этой области проблем и, в частности, в том, чтобы показать адекватность методов распознавания задаче представления больших массивов информации в обозримом виде. Эта последняя задача важна для разработки новых средств интерпретация результатов массовых обследований, для усовершенствования самой организации массовых обследований и для создания эффективных средств сжатия информации с целью ее хранения в ЦВМ.

Можно выделить три основных типа задач, решаемых методами распознавания образов:

1. задача обучения распознаванию образов с учителем (или выработки и использования решающего правила);

2. задача обучения распознаванию образов без учителя (или, что то же самое, задача автоматической классификации, таксономии и т.п.);

3. задача разделения параметров на группы "близких".

Предложено много алгоритмов решения этих задач. С помощью предложенных алгоритмов решались различные прикладные вопросы из области техники, экономики, биологии, медицины и др. В некоторых случаях одни и те же задачи решались с помощью разных алгоритмов. Проведенные исследования привели многих специалистов к заключению, что изученные алгоритмы имеют примерно одну эффективность и что основные факторы, обеспечивающие успех решения, связаны, как правило, с удачной постановкой задачи. По этой причине в статье основное внимание уделено постановкам задач. Однако больше всего места в работе уделено вопросу, как при постановке задачи формулируются условия того, что считать успешным решени-

ем и что рассматривать как неудачу. Это самый трудный вопрос применения методов распознавания образов к социологическим задачам.

Отчасти в связи со всем выше сказанным, а в основном в связи с характером подачи материала, отражающим лишь методическую сторону проблемы применения методов распознавания образов в социологии, в работе решено не цитировать другие работы, хотя это можно было бы сделать во многих местах статьи.

В статье имеется два раздела. Первый раздел посвящен описанию основных задач, решаемых методами распознавания образов. Во втором, главном разделе сначала рассмотрены общие подходы к оценке результатов, получаемых при решении задач методами распознавания. После этого дается описание двух новых задач: первая задача связана с разработкой машинных методов организации "малой" выборки, аппроксимирующей "большое" обследование; вторая задача связана с разработкой машинных методов выделения "маленького" вопросника, аппроксимирующего "большую" анкету. Эти задачи даны не столько как примеры применения методов распознавания, сколько с целью показать как можно строить формальные критерии оценки качества получаемых с помощью методов распознавания результатов. В конце второго раздела описывается подход к решению общей задачи компактного представления данных большого обследования - подход, связанный с выработкой языка "качественного" (краткого) описания сложных объектов.

## I. ОСНОВНЫЕ ЗАДАЧИ

В этом разделе будут описаны три основные задачи применения методов распознаванию образов, перечисленные во введении.

Задача выработки и использования решающего правила (обучения распознаванию образов с учителем). При решении этой задачи предполагается, что имеется информация о небольшой группе объектов, взятых из нескольких классов и для каждого объекта известна его принадлежность к классам.

Например, это могут быть результаты клинических анализов больных воспалением легких и раком легких или информация о технологических процессах, отличающихся сортом выходного продукта. Это может быть информация о результатах обследования условий жизни сельских жителей, о которых потом стало известно, кто из них остался в селе, а кто уехал из села в город; это может быть информация о структуре свободного времени семей, о которых потом стало известно, какая из них сохранилась, а какая распалась и т.д. и т.п. Проблемя заключается в том, чтобы определить как по результатам анализов поставить дифференциальный диагноз больного ( у него воспаление легких или рак легких) или как по результатам обследования структуры свободного времени семей оценить вероятность того, что семья не распадается.

При решении такого рода задач возникают два вопроса : как вырабатывать решающее правило для отнесения объектов к классам и как использовать это правило для распознавания принадлежности к классу большого числа новых объектов.

При решении этих вопросов объект, который задан набором параметров (признаков, характеристик, ответов и т.п.), рассматривается как точка (вектор) в некотором пространстве $X$. Значения координат точки откладываются на осях, каждая из которых соответствует определенному параметру $x_i$.

В рамках таких геометрических представлений "близким" объектам соответствует множество "близких" в введенном пространстве $X$ точек, и поэтому каждому классу объектов соответствует своя область этого пространства. Таким образом, правило отнесения объекта к некоторому классу есть способ указать область пространства $X$, которая соответствует этому классу.

Если, например, в пространстве $X$ задана система сфер, каждая из которых охватывает область, соответствующую определенному классу объектов, то правило отнесения объекта к классу сводится к определению того, внутри какой сферы расположена соответствующая искомому объекту точка. Это правило можно записать следующим образом. Пусть $\{x_i, i=1,2,\dots,k\}$

есть множество центров сфер, каждая из которых охватывает область одного из $k$ классов объектов; пусть $\{R_i,\ i=1,2,\dots,k\}$ - радиусы соответствующих сфер. Тогда точка $x$ относится к классу $\rho$, если выполняются два условия:

$$\left.\begin{aligned} &\min\{r(x,z_i)\} = r(x,z_\rho) \\ &r(x,z_\rho) \le R_\rho \end{aligned}\right\} \qquad (1)$$

где $r(x,z_i)$ - расстояние между точкой $x$ и точкой $z_i$ в пространстве $X$.

Другой способ определения класса, к которому относится объект $x$, состоит в нахождении индекса $\rho$, который удовлетворяет условию

$$\max\{(c_i,x)\} = (c_\rho,x), \qquad (2)$$

где $(w,v)$ - скалярное произведение векторов $w$ и $v$ (аналогично расстоянию в предыдущем правиле в этом правиле выступает скалярное произведение: оно есть функция "близости" между парой точек); $\{c_i,\ i=1,2,\dots,k\}$ - "опорные" точки. Второе правило, таким образом, определяет, что опорная точка, к которой искомая точка $x$ "ближе" в смысле скалярного произведения и определяет индекс класса, к которому принадлежит соответствующей этой точке $x$ объект.

Задача выработки правила отнесения заключается в том, чтобы с помощью примеров точек из известных классов найти центры, охватывающих эти классы сфер или опорные точки этих классов или другую систему параметров, с помощью которой можно задать систему функций "близости" точки к классу.

Простейший способ решения такой задачи следующий. Из известных примеров выбираются точки, соответствующие одному классу. Находится их центр тяжести, т.е. точка, значение координат которой есть среднеарифметические значения соответствующих координат выбранных точек. Этот центр тяжести выбирается за "опорную" точку этого класса. Аналогично определяются "опорные" точки других классов. После чего искомое решающее правило дается условием (2).

В тех случаях, когда возникает задача выработки решающего правила отнесения объекта к классу, легко организовать

испытание качества выработанного правила. Для этого оставляется часть множества объектов, для которых точно известна их принадлежность к классам. На этих объектах определяется их принадлежность к классам с помощью выработанного правила. Процент правильно отнесенных с помощью выработанного правила объектов служит мерой его качества. Если мера качества высокая, то правило можно использовать. В противном случае необходимо или искать другое правило или дополнить описание рассматриваемых объектов дополнительными признаками или увеличить множество объектов, для которых известны их принадлежность к классам и которые служат основой для выработки правила.

Возможность строго подойти к оценке результата обеспечила ту легкость, с которой задача обучения распознования образов с учителем стала применяться в прикладных задачах. Эта возможность, кроме того, открыла путь целенаправленного приспособления общего метода к решению конкретной задачи.

Задача автоматической классификации (обучения распознаванию образов без учителя). Эта задача хорошо знакома тем, кто занимается вопросам обработки больших массивов информации. Она ставится следующим образом.

Имеется множество наборов параметров, каждый из которых характеризует отдельный объект. Требуется найти разбиение их на "однородные" группы, т.е. такое разбиение, при котором внутри выделенных групп оказались бы "близкие" наборы. Такая постановка задачи соответствует нашим интуитивным представлениям о классификации считать элементами одного класса скопления "близки" объектов.

Естественно, что алгоритмы, решающие эту задачу, дают одновременно и границу областей пространства наборов на области, в каждой из которых находится своя "однородная" группа. Таким образом, алгоритм решения этой задачи дает решающее правило, по которому можно узнавать принадлежность новых объектов к классам в рамках сформированной классификации.

Из сказанного ясно, что задача автоматической классификации "похожа" на ранее описанную задачу обучения распознаванию образов с учителем. И хотя в постановке задачи автоматической классификации не предполагается наличие примеров наборов параметров, для которых известна априорная классификация, и задача заключается именно в отыскании классификации, вместе с тем, подходы к той и другой задаче, - общие. Они заключаются в представлении о том,что в одном классе должны быть "близкие" между собой объекты. По этой причине и алгоритмы решения второй задачи весьма близки по общей идее к алгоритмам решения первой задачи. Отличие новой задачи хорошо отражает другое ее название - "задача обучения распознаванию образов без учителя".

При постановке задачи об автоматической классификации в некоторых случаях заранее указывается на сколько групп необходимо разделить искомое множество объектов. В этом случае задачу удается естественно сформулировать как задачу на поиск "лучшего" разбиения множества на группы, т.е. такого разбиения, при котором получаются наиболее "плотные" группы. При этом заранее нельзя задаться степенью "плотности" групп. Наоборот, при другом подходе заранее требуется выделить группы, у которых "плотность" будет не ниже некоторой заданной. В этом случае зато можно поставить задачу о нахождении разбиения множества на группы с "плотностью" не ниже заданной так, чтобы при этом число групп было наименьшим из возможных.

Задача группировки параметров.Данные, полученные в результате обследования можно представить в виде прямоугольной матрицы, строки которой соответствуют объектам,а столбцы - параметрам. В терминах этого представления рассмотренная выше задача автоматической классификации заключается в сортировке строк матрицы данных на группы "близких".

Если теперь повернуть матрицу данных так, чтобы строки соответствовали параметрам, а столбцы - объектам и применить формально имеющиеся методы автоматической классификации для сортировки строк повернутой матрицы, то это будет уже решение новой задачи разделения параметров на "плотные"

группы. При подходе к этой задаче с содержательной стороны, необходимо иметь в виду, что параметру обычно можно приписать знак и что параметры, которые отличаются лишь знаком характеризуют одну величину. Учет этого обстоятельства легко осуществить специальным выбором исходной функции "близости" между параметрами. В остальном же для решения задачи разделения параметров можно применять алгоритмы автоматической классификации. Целью разделения параметров является получение таких групп, каждую из которых можно рассматривать как набор косвенных величин, отражающих влияние "внутреннего" фактора.

В специальном случае, когда в качестве исходной меры "близости" между двумя параметрами выбирается коэффициент корреляции между ними, удается строго сформулировать задачу о нахождении "лучшего" разбиения параметров на группы. Эта задача, как показали исследования, оказалась тесно связанной с задачами факторного анализа.

## 2. КРИТЕРИИ ОЦЕНКИ КАЧЕСТВА РЕШЕНИЯ ЗАДАЧ МЕТОДАМИ РАСПОЗНАВАНИЯ ОБРАЗОВ

В первом разделе уже говорилось, что задача выработки и использования решающего правила имеет простой способ оценки качества решения. Качество оценивается по тому, насколько результаты отнесения объектов к классам с помощью выработанного правила совпадают с действительной принадлежностью объектов к классам. Конечно, для того, чтобы уметь применить этот способ, необходимо иметь достаточно много объектов с известной принадлежностью к классам. Часто однако имеется недостаточная статистика. Тогда возникает вопрос, как априорный материал лучшим образом поделить на две части так, чтобы одна часть была использована для выработки решающего правила, а другая – для проведения испытания правила, т.е. для получения оценки качества построенного правила. В таких случаях иногда используют режим "скользящего" распознавания. При этом выработка решающего правила проводится многократно. Каждый раз исход-

ный материал делят на две не равные части: почти весь материал используют для выработки правила, а на оставшихся одном - двух объектах, не использованных при выработке правила, проверяют правило.

После этого исходный материал снова разбивают на две не равные части, причем так, чтобы один - два объекта, которые оставляются для испытаний нового правила были отличны от тех, которые были использованы для испытания предыдущего правила. Опять проводится построение правила и его испытание. Так делается до тех пор, пока в испытаниях не побывает весь исходный материал. Процент правильных ответов на всех испытаниях принимается за оценку качества выработанного правила.

Таким образом, при решении первой задачи ясно как оценить успешность полученного решения.

Совершенно иначе дела обстоят с решениями задач второго и третьего типа. В некотором смысле постановка этих задач даже противоречива: требуется отыскать "хорошую" сортировку объектов или параметров, для которых неизвестна их классификация, но определить насколько "хорошей" получилась сортировка можно лишь в случае, когда заранее задана классификация.

При решении этого типа задач исходят обычно из следующего общего методического принципа. Рассмотрим ряд разных практических примеров, когда известна установившаяся классификация. Попробуем в этих случаях решить задачи автоматической классификации и группировки параметров. Если оказывается, что получается хорошее совпадение сформированной классификации и установившейся ( сформированной группировки параметров и установившейся), то это служит основанием использовать методы решения этих задач для получения классификации объектов и в неизвестной ситуации.

Есть другая возможность получить оценки качества решения задач автоматической классификации и группировки параметров. Она заключается в том, чтобы такие задачи не решать изолированно, а "помещать" их в более общую задачу,

чтобы рассматривать решение таких задач как этапы в решении "большой" задачи, для которой известно, что такое хорошее решение и что такое плохое решение. В такой ситуации оценка качества решения задач автоматической классификации и группировки параметров естественно проводится исходя из того, насколько хорошим получается решение "охватывающей" целевой задачи.

Ниже описываются два примера целевых задач, в которых задачи автоматической классификации и группировки параметров выступают как вспомогательные.

Задача о "выборке". Эта задача состоит в организации способа отбора небольшой группы людей (выборки), по ответам которых можно с уверенностью судить об ответах интересующей большой совокупности людей. Решение ее дает ответы на вопросы - как найти тех людей, которые будут опрашиваться.

В случаях, когда мы можем часто проводить "обширные" обследования, задача о выборке не стоит так остро. Имеющиеся общие статистические данные как правило бывают достаточны для того, чтобы хорошо провести обследование. В этом случае "внутри", в рамках данных самого обследования, оказывается возможным "проверить" устойчивость оцениваемых средних или корректно "исправить" полученную статистику (выбросить, например, некоторые наблюдения), так,что после исправления можно определить устойчивые оценки средних. Способы проверки и исправления полученных данных в "обширных" обследованиях можно считать разработанными.

Задача о "выборке" возникает тогда, когда надо создать службу такого оперативного сбора информации, при котором "обширные" обследования есть возможность проводить только изредка. В этих случаях необходимо уметь отыскать небольшую группу "типичных" для изучаемой совокупности людей.

Будем исходить из представлений, что в отношении исследуемого вопроса интересующая совокупность людей состоит из небольшого числа четко различающихся групп: внутри группы люди "почти одинаково" отвечают на поставленные вопросы,

а между ответами членов разных групп есть "большие" различия. Конечно, люди могут со временем сильно изменить свои установки, но мы предполагаем еще, что эти возможные изменения, хотя и могут происходить очень быстро во времени, у большинства членов одной группы происходят "почти одновременно". Предположение, таким образом, заключается в том, что люди разделяются на сильно различающиеся группы по установкам и что принадлежность к группе есть намного более стабильный фактор, чем фактическое изменение конкретной установки.

Основываясь на этих предположениях, можно предложить следующую принципиальную схему построения выборки.

Возьмем данные одного обширного обследования, которое редко, но проводить все есть возможность. Эти данные, как отмечалось, образуют массив ответов представительной выборки. Пусть ответы одного ( $i$ -ого) человека задаются в этой выборке вектором $x_i = \{x_i^{(1)}, \dots, x_i^{(n)}\}$ где $n$ - - число ответов, а $x_i^{(j)}$ - значение $i$ -ого человека на $j$ -ый вопрос. Пусть всего в рассматриваемой представительной выборке имеется $N$ человек. Будем интересоваться тем, как построить выборку из $m$ ($m << N$) человек такую, чтобы выбор средних от значений ответов по этой выборке $\bar{x}_{m,i}$ "мало" отличался от набора средних значений по ответам представительной выборки $\bar{x}_{n,i}$ . Сформулируем это требование в виде формального критерия найти выборку из $m$ человек, удовлетворяющую условию

$$J = \sqrt{\frac{(\bar{x}_{m,i} - \bar{x}_{n,i})^2}{(\bar{x}_{n,i})^2}} < \delta \qquad (3)$$

где $\delta$ - наперед заданное число, такое, что $1 > \delta > 0$.

Для того, чтобы найти такую выборку, методами автоматической классификации разделим представительную выборку на $K$ групп людей. Отберем случайным образом из каждой полученной $s$-ой группы $m \cdot \frac{N_s}{N}$ членов ( $N_s$ -общее число представителей в $s$ -ой группе). Для отобранных $m$ людей

подсчитаем вектор средних значений по ответам этих людей $\bar{x}_m$. Если подставив в (3) $\bar{x}_m$ вместо $\bar{x}_{m,\ell}$, получим $J < \delta$, то отобранную группу и примем за искомую выборку. В противном случае, предпримем новую классификацию представительной выборки на другое число групп. Если в результате нескольких таких проб не удастся найти выборку, дающую $J < \delta$, то несколько увеличим $\delta$ и снова проделаем несколько проб отбора $m$ человек по способу, описанному выше. Так будем делать до тех пор, пока не найдем такое $\delta^*$, такую классификацию и такую выборку, что $J < \delta^*$. Тем самым мы не только получим выборку, но и будем знать грубую оценку "качества" этой выборки $\delta^*$ и число групп "близких" по ответам людей, на которое разбивается представительная выборка. Полученная выборка используется в оперативной работе до тех пор, пока не проводится следующее "обширное" обследование. Последнее дает возможность скорректировать эту выборку или заменить ее новой.

Описанная схема не должна рассматриваться как алгоритм поиска выборки. Она характеризует лишь основную идею такого алгоритма, и, в частности, объясняет как при построении такого способа "работает" метод автоматической классификации.

Для того, чтобы как-то обосновать целый ряд содержательных предположений описанной выше схемы организации выборки, необходимо вообще говоря иметь материалы нескольких подряд проведенных "больших" обследований. Однако и в случае данных одного такого обследования есть некоторая возможность исследовать эту схему. Для этого надо, чтобы анкета такого обследования была сильно избыточна. В этом случае разделим эту анкету на поданкеты. Сделаем это исходя из представлений о том, что все поданкеты должны характеризовать одни и те же "факторы" в одинаковой степени. Данные по каждой поданкете можно рассматривать как "самостоятельно" проведенное обследование. Если одно из этих подследований по поданкете можно принять за базовое и на его основе построить выборку, то обследования по другим поданкетам можно будет использовать для проверки сделанной выборки.

Задача о "мини" - анкете. Имеется большая анкета для регулярного наблюдения за некоторой определенной совокупностью людей. Пусть эта анкета составлена из вопросов, каждый из которых "хорошо отражает" лишь один из факторов. Пусть изучаемых факторов будет всего несколько, хотя вопросов в анкете может быть несколько десятков и даже сотен. Цель наблюдения - регулярно давать средние значения по ответам изучаемой совокупности людей на все вопросы большой анкеты.

Предположим, что хотя со временем отдельный человек может изменить свои ответы, но эти изменения носят "групповой" характер - человек "на самом деле" изменяет свои реакции на факторы, отраженные в вопросах анкеты. Поэтому как только изменятся его реакция на некоторый фактор, так сразу же изменятся соответствующим образом его ответы на все связанные с этим фактором вопросы. В то же время ответы на вопросы, связанные с другим фактором, сохранятся.

Эти предположения делают почти очевидным схему сокращения анкеты. Рассмотрим однократное обследование искомой совокупности людей по всей большой анкете. На основе полученных с ее помощью ответов построим матрицу корреляций между всеми вопросами. С помощью алгоритмов группировки параметров (вопросов) выделим основные факторы. Определим, какие из вопросов с каким фактором связаны. Из групп вопросов каждого фактора выберем 2 - 3 вопроса так, чтобы ответы на эти два-три вопроса давали хорошую оценку значения фактора, которое можно вычислить точно, зная ответы на все вопросы, относящиеся к фактору.

Пусть для простоты каждый из вопросов анкеты требует ответов только "да" и "нет", которые кодируются +I и -I соответственно. Для данного конкретного человека относительно его ответов на каждый вопрос запоминается, совпадает или не совпадает знак ответа на этот вопрос со знаком соответствующего ему фактора.

Рассмотрим теперь "выборку" вопросов по два-три от каждого фактора, о которой говорилось выше. Проведем с помощью этой "выборки" новый опрос обследуемого человека. Вычислим факторы для данного человека по новому опросу . По предположению вопросы из "выборки" дают возможность провести эти вычисления достаточно точно. После этого легко по знакам факторов и таблице совпадений знаков факторов со знаками вопросов большой анкеты "восстановить" возможные ответы этого человека на не участвовавшие в "выборке" вопросы большой анкеты в новом опросе. Это можно сделать для каждого человека искомой совокупности людей.

Если имеется два подряд проведенных обследования на этой совокупности людей по всей большой анкете, то легко построить проверку качества сделанного малого вопросника, оценивая результаты просто по числу угаданных ответов.Если это число является неудовлетворительным, то можно изменить вопросник, заново применив методы группировки параметров с разделением вопросов большой анкеты на другое число групп факторов. Кроме того, всегда есть возможность, увеличивая число вопросов одного фактора, которые отбираются в малый вопросник обеспечить требуемый уровень правильно получаемых ответов. Тот малый вопросник, который обеспечивает этот уровень и выбирается в качестве "мини"-анкеты.

Задача компактного описания матрицы данных. Несколько раз уже отмечалось, что основной целью применения методов распознаванию образов в социологии является задача компактного представления больших массивов эмпирического материала. Пусть эмпирический материал имеет вид матрицы данных. Если с помощью методов группировки параметров разделить столбцы матрицы, то получим группы вопросов, которые (эти группы) легко интерпретировать как вопросники, связанные каждый с одним содержательно интерпретируемым фактором. Важно, что методы группировки параметров дают помимо группировки вопросов возможность вычислять значения факторов для отдельного человека (ввести факторные шкалы). Тем самым большой набор ответов на косвенные вопросы можно заме-

нить небольшим набором чисел, представляющих из себя значения хорошо интерпретируемых факторов. После того, как произведено разделение вопросов на группы, в каждой группе отдельно можно разбить людей на типы методами автоматической классификации. Каждое такое разбиение дает типологию обследованных людей по своему фактору. Таким образом, для каждого человека будет построено описание, характеризующее к какому типу по какому фактору относится данный индивид. Более того, каждое отдельное описание можно будет конкретизировать, указав величины значений факторов "типичных" для индивидов, характеризуемых этим описанием. Такое представление матрицы данных естественно рассматривать как язык описания. Словарем этого языка являются названия факторов и типов людей по каждому фактору. С помощью этого языка можно строить описания новых опрошенных по той же анкете людей. Ясно, что среди получаемых таким образом описаний могут быть как те, которые уже имеются в исходной матрице данных, так и новые . Если в таком процессе составления описаний новых людей, новых описаний будет мало, то естественно считать полученное представление искомой матрицы данных "устойчивым". В противном случае, надо подбирать другой язык. В частности, по другому (на другое число), разделяя вопросы на группы и людей на типы, можно получить другой словарь языка. Тем самым будет получено новое компактное представление искомой матрицы данных, которой аналогично предыдущему может быть проверено на "устойчивость".

## KUJUNDITE ERISTAMISE MEETODID SOTSIOLOOGIAS

I.B. Mutšnik
(Moskva)

Resümee

Artiklis antakse ülevaade kujundite eristamise meetodite põhilistest rakendusvõimalustest sotsioloogias. Ülesanded jaotatakse kolme olulisesse tüüpi.

1. Kujundite eristamine "õpetajaga" (otsustuseeskirja väljatöötamine).

2. Kujundite eristamine "õpetajata" (automaatne klassifitseerimine, taksonoomia).

3. Parameetrite rühmitamine mingis mõttes läheduse järgi.

Tutvustatakse ka meetodeid ülesande lahendi headuse hindamiseks, aga samuti rida konkreetseid ülesandeid, mis on ülaltoodud probleemistikuga tihedasti seotud: representatiivse väljavõtte valimine, "mini"-ankeedi koostamine, lähteandmete maatriksi kompaktne üleskirjutamine.

## METHODS OF PATTERN RECOGNITION IN SOCIOLOGY

I.B. Mutohnik
(Moscow)

Summary

The review of the main possibilities for application of pattern recognition methods in sociology is given. The problems are divided into three types.

1. Pattern recognition with the "teacher" (the working out of rules for decision-making).

2. Pattern recognition without the "teacher" (automatic classification, taxonomy).

3. The classification of parameters on the basis of their proximity (in some sense).

Some methods of estimating the goodness of the solution of a problem are introduced. Some practical issues closely connected with the afore-mentioned problems are dealt with such as the representative sampling, the construction of a 'mini' questionnaire, the compact reproduction of a data matrix.

# ОБ ОДНОЙ МАТЕМАТИЧЕСКОЙ ФОРМАЛИЗАЦИИ ЗАДАЧ РАСПОЗНАВАНИЯ ОБРАЗОВ

Э.Тийт
(Тарту)

Имеется много разных попыток рассмотреть все задачи распознавания образов с единой точки зрения, и дать соответственно их классификацию. Настоящая заметка является попыткой математически формализовать процесс распознавания образов как процесс принятия решений, и классифицировать возникающие при этом задачи.

I. Формулировка задачи

Общий процесс распознавания образов может быть описан следующим образом:

I)имеется некоторое пространство $R$ ( пространство "точек","объектов","индивидов"),элемент которого обозначается буквой $x$ . Размерность $r$ пространства $R$ может быть конечной ( $r \geq 1$), бесконечной счетной или несчетной.

Предполагается, что в пространстве $R$ определена некоторая метрика, которая сопоставляет двум элементам $x$ и $y$ с $R$ их расстояние $d(x, y)$ (действительное число). Притом это расстояние может быть обобщенным; не обязательно требовать выполнения всех аксиом расстояния; аксиоматика может быть более слабой.

2) Имеется некоторое пространство $A$ (пространство "образов","групп","таксонов"), элементы которого обозначаются буквами $A_i$ ( $i = 1, 2, \ldots$ ).Число элементов $A_i$ может быть конечным или счетным; размерность $a$ пространства $A$ также конечным или счетным,притом часто $a = 1$.

3) имеется некоторая случайная функция $X_x$, $x \in R$

со значениями в пространстве $A$ : для каждого элемента $x$ определяется случайная величина $X_x$ такая, что

$$P(X_x = A_i) = p_i \qquad \sum_{A_i \in A} p_i = 1 \tag{I}$$

В частном случае случайные величины $X_k$ могут быть вырожденными: для каждой точки $k \in R$ существует однозначно определенный элемент $A_i \in A$ такой, что

$$P(X_k = A_i) = 1,$$
$$P(X_k = A_k) = 0. \qquad \text{если } i = k$$

В таком случае можно говорить и о детерминированном отображении $\varphi(R \to A)$ ; для каждого $x \in R$ определяется

$$\varphi(x) = A_i . \qquad (2)$$

Проинтерпретируем теперь введенные понятия.

$R$ – множество всевозможных (но не объективно существующих) объектов. Размерность пространства $R$ определяется множеством признаков, характеризующих объекты. Это множество может быть счетным (последовательно рассматриваемые психологические или биологические признаки) или несчетным (точки изображений, кривых). Почти всегда известным является только конечное множество из них; значит, практически объекты рассматриваются в $p$ –мерном подпространстве пространства $R$ , где $p$ (фиксированное или нефиксированное) конечное число.

$A$ – множество всевозможных образов.

Практически представляет интерес только конечное множество из всевозможных точек $A_i$ .

Для точек $A_i$ возможны разные интерпретации. Самым известным является представление точек $A$ как подмножеств пространства $R$ : существует некоторое разбиение множества $R$ на непересекающиеся части $S_i$,

$$S_i \cap S_j = \Phi, \qquad (3)$$

$$\cup S_i = R . \qquad (4)$$

Тогда можно считать множества $S_i \subset R$ эквивалентными точкам $A_i \in A$ и соответственно с соотношением (2):

$$x \in S_i$$

Более общее представление можно получить, избегая условия (3) и (4), и рассматривая случайное отображение; равенство (I) в таком случае эквивалентно утверждению

$$P(x \in S_i) = p_i .$$

Еще более общее содержание для точек $A_i$ дает их рассмотрение как решений. Соотношение (2) означает в таком случае, что относительно точки $x$ принимается некоторое $i$ -тое решение (таким может быть утверждение, что $x$ принадлежит к некоторой группе $i$ или о точке $x$ невозможно сказать к какой группе она принадлежит , и т.п.).

Если задача классификации однозначна (каждая точка принадлежит только в один образ или дается распределение вероятности для непересекающихся образов), то целесообразно рассмотреть пространство $A$ как одномерное. В случае же, если желательно получить частично пересекающуюся классификацию (обычно говорится, что для каждого класса определены подклассы нескольких уровней), то пространство $A$ приходится рассматривать как $k$-мерное. Такая классификация в формальном виде называется линнеевской.

Отображения $X_x$ , $\varphi(x)$ - это решающие функции. В случае однозначного соответствия (2) получим детерминированную, в общем же случае (I) - стохастическую процедуру.

В зависимости от того, какие элементы из соответствия

$$R \xrightarrow{X_k} A \qquad (5)$$

или

$$R \xrightarrow{\varphi} A \qquad (6)$$

неизвестны или известны неполностью, можно говорить о нескольких типах задач распознавания образов.

2. Распознавание при заданных образах

Предположим, что пространства $R$ и $A$ известны, но неизвестна (случайная или детерминированная) функция $X_x$ . Эта задача практически реализуется обычно таким образом,что

для некоторого (конечного) подмножества $K \in R$ известны значения функции $X_k$ (или $\varphi(x)$).

Если предполагается, что существует детерминированная решающая функция, то ее нахождение является просто задачей аппроксимации функции, в частном случае задачей экстраполяции или интерполяции.

В случае, когда точкам $A_i$ однозначно соответствуют подмножества $S_i$ пространства $R$, то нахождение решающих функций приводит к нахождению т.н. дискриминирующих функций, которые разделяют пространство $R$ в непересекающиеся части. Конечно, такие дискриминирующие функции не определяются однозначно на основании конечного числа заданных точек, и для конкретизации задачи:

1) задается конкретная форма функции (линейная, кусочно-линейная, параболическая и т.д.);

2) определяются некоторые критерии оптимальности.

Одна возможность оптимизации - экзаминация. Для этого заданное множество $K$ разбивается случайным образом на две части $K_1$ и $K_2$ (удовлетворяющие условиям (3) и (4) или только (4). Решающие функции строятся на основании множества $K_1$ (обучающее множество), а затем на множестве $K_2$, определяют эмпирические вероятности ошибок $L_{i/j}$ (вероятность отнесения элемента в образ $i$, если он принадлежит к образу $j$).

Иногда при оптимизации требуется учитывать штрафы $c_{i/j}$ за ошибки; такие решающие правила называются байесовскими.

При определении стохастической решающей процедуры множество $K$ разбивается на части $K_i$ по следующему правилу.

Если реализация функции $X_x$ в точке $x$ равняется $A_i$, то $x \in K_i$

В результате получаются выборки

$$\{x \cdot x \in K_i\} \qquad (7)$$

на основании которых определяются эмпирические распределения $P_i$ . Эта задача решается сравнительно просто в случае, когда общий тип распределения считается априорно известным, и только параметры нуждаются в конкретизации; последние оцениваются по выборке, а также целесообразно проверять согласие эмпирического и априорного распределения.

Решающие функции в таком случае часто определяются методом максимального правдоподобия:

$$x \in A_i \text{ , если } f_i(x) = \max f_j(x)$$

где $f_i$ – эмпирическая плотность выборки (7).

Если учесть и априорное распределение (например частоты элементов в группах $K_i$ (см. (7)), то решающее правило можно определить с помощью формулы Бейеса:

$$x \in A_i, \text{ если } p_i f_i(x) = \max p_j f_j(x)$$

или при учитывании штрафов $c_{j/i}$ :

$$x \in A_i, \text{ если } \sum_j p_j f_j(x)\, c_{j/i} = \min_k p_j f_j(x) c_{j/k}. \qquad (8)$$

Заметим, что в некоторых случаях детерминированные и стохастические решения формально совпадают – если распределения случайных величин $y_i$ относится к типу экспоненциальных, то решающие функции, полученные из формулы (8), являются линейными, если показатель в формуле $f_i(x)$ в первой степени, и квадратичными, если показатель квадратичный. В частном случае, при нормальном распределении, получаются линейные дискриминирующие функции, если ковариационные матрицы распределений $f_i$ совпадают.

## 3. Таксономия

Второй тип задач такой, при котором неизвестно пространство $A$ ; эта задача решима в случае , когда точки $A_i$ считаются эквивалентными множествам $S_i$ из пространства $R$ .

В данном случае необходимо, чтобы в пространстве была бы определена некоторая "мера компактности", которая в общем случае является функцией распределения (этой мерой может быть и лебеговая мера в пространстве $R$ ; тогда дополнительного определения не надо).

Обычно для решения задачи второго типа задано множество $K \subset R$, для каждой точки $x \in R$ требуется определить функцию $\varphi(x)=A$ или $X_x = A$ . Если эта часть задачи решена, то для остальных множеств $K_i \subset R$ задача распознавания образов решается аналогично предыдущему случаю.

Большое количество методов решения задач второго типа (т.н. задачи таксономии) можно разделить на 2 класса: прямые и итеративные. Прямые методы используют сразу всю информацию о расположении элементов $x_i \in K$ в пространстве $R$ и дают сразу окончательное описание функций $\varphi(x)$ или $X_x$ . Это описание получается индуктивным или дедуктивным методом, обычно иерархически. Например, в индуктивной процедуре начинают с объединения в одну группу $A_i$ двух самых близких друг к другу элементов $x_1'$ и $x_2'$ из множества $K$ ; затем объединяют в группу еще некоторое количество элементов, исходя из определенной меры компактности ; окончание образования группы определяется специальным "правилом остановки". Затем начинается образование второй группы и т.д. Возможны и правила индуктивных процедур, включающие элементы переменно в разные группы. При дедуктивных методах множество $K$ делится на две части, затем полученные множества делятся на части и т.д.; мерой компактности может быть, например, отношение межгруппового и внутригруппового среднего расстояния.

Описанные методы применимы в случае не очень многочисленных множеств $K$ . Если же численность множества $K$ большая, то более целесообразно применить некоторые итерационные методы, которые на начальном этапе не применяют всю известную информацию; принятые решения корректируются итерационным путем, а тем самым открывается и возможность судить об их качестве. К сожалению, для большинства итерационных методов не доказана сходимость.

Итерационными являются как детерминированные, так и стохастические методы распознавания образов. В первом случае в ходе итерации уточняются дискриминирующие функции, во втором случае - параметры распределения, и по ним - дискриминирующие функции.

Задачи второго типа можно рассмотреть как комплексные, состоящие из двух частей - определение образов (таксонов) $A_1, A_2, \ldots$ на материале $\mathcal{K}$ и определения решающей функции $\varphi(x)$ или $\chi_{\mathcal{K}}$ - эта задача первого типа. Но часто ограничиваются и только первой половиной задачи - группируется только множество $\mathcal{K}$ и решающих правил для дальнейшей классификации множества $R$ не строят.

## 4. Определение информативных признаков

Третий тип задач связан с пространством $R$. Сюда относятся главным образом задачи выделения т.н. "информативного подмножества признаков". Особенно в последнее время такие задачи стали в центре внимания исследователей. С этой задачей связано и определение метрики, компактности и других функций в пространстве $R$, а также упорядочение и определение весов для координатных осей (признаков).

Задачи третьего типа обычно не решаются отдельно, а в связи с задачами двух остальных типов, особенно со вторым типом.

Самыми комплексными являются проблемы, содержащие задачи всех трех типов (или элементы этих задач).

## 5. Историческое развитие распознавания образов

Об историческом развитии методов распознавания образов следует отметить, что хотя стохастические решения в общем являются более сложными, все-таки первые решения задач, принадлежащих к области распознавания образов статистическим анализом являются именно стохастическими.

Сюда относится в первую очередь классический дискриминантный анализ фишера, являющийся решением стохастической

задачи первого типа для двух нормально распределенных случайных величин (выборки $K_1$ и $K_2$, имеющих равную ковариационную матрицу. Близким к этому методу является и $T^2$ - тест для проверки многомерных гипотез, а также методика, связанная с расстоянием Махаланобиса. С другой стороны относятся сюда некоторые результаты теории выборки (т.н.стратифицированные выборки).

С началом применения электронно-вычислительной техники возникли серьезные вопросы, связанные с познаванием текста, фигур и т.д. Бурно начали развиваться детерминированные методы распознавания образов при помощи обучающих выборок.

В то же время в целях биологической классификации в основном биологами и биометристами-статистиками были развиты и методы таксономии (задачи второго типа). Обе эти методики на первых порах рассматривались преимущественно детерминированными и они развивались раздельно.

В течение последнего десятилетия начались попытки найти более точные, стохастические (рандомизированные) решения как для задач первого, так и для задач второго типа. Найдены и общие черты во всех рассматриваемых задачах.

В то же время дойдено до вывода, что всякие результаты группировок и распознавания образов сильно зависят от набора признаков, на основании которых ведется распознавание. Так как набор признаков часто определялся субъективно исследователем, можно всякую классификацию считать в некоторой мере субъективной. Для преодоления этого недостатка возникли в последнее время исследования для решения задач третьего типа.

## ÜHEST KUJUNDITE ERISTAMISE ÜLESANNETE MATEMAATILISEST FORMALISEERIMISEST

E. Tiit
(Tartu)

Resümee

Käesolevas artiklis esitatakse kujundite eristamise protsess teatava otsustusprotsessina, ning antakse vastavalt sellele esinevate ülesandetüüpide klassifikatsioon.

Kõigi kujundite eristamise ülesannete lahendamise tulemusena saame järgmised matemaatilised objektid.

1) Punktide või objektide x ruum R , mille dimensioon võib olla lõplik või ka lõpmatu.

Eeldatakse,et ruumis R on defineeritud 2 punkti vaheline kaugus ( või üldistatud kaugus) d (x,y);

2) Kujundite ehk rühmade ruum A elementidega $A_i$, rühmade arv, samuti ka ruumi dimensioon on ülimalt loenduv.

3) Juhuslik funktsioon $\chi_x$, $x \in R$ väärtustega ruumis A: iga elemendi x jaoks määratakse juhuslik suurus $\chi_x$ nii, et

$$P(\chi_x = A_i) = p_i$$

$$\sum_{A_i \in A} p_i = 1$$

Erijuhul,kui $\chi_x$ on kõdunud , on meil tegemist determineeritud kujutisega $\varphi$ ( R $\to$ A).

Kui R ja A on tuntud,kuid $\chi_x$ on määramata, on meil tegemist õpperühmaga kujundite eristamise ülesandega.

Kui R on tuntud, A aga tundmata, on tegemist taksonoomia ülesannetega.

Kui täpsustamist vajab ka ruum R, on tarvis lahendada kõigepealt tunnuste informatiivse alamhulga määramise küsimus.

Lõpuks esitatakse lühiülevaade nimetatud ülesannete lahendamismeetodite ajaloolisest arengust.

# MATHEMATICAL FORMALIZATION OF PATTERN RECOGNITION PROBLEMS

E. Tiit
(Tartu)

Summary

In the present paper the process of pattern recognition is described as a decision-making process; the classification of several types of pattern recognition problems is given.

While solving some pattern recognition problems we get the following mathematical objects:

1. The points' or objects' room R. The dimension of room R is finite or infinite. We suggest that the distance (or the generalized distance) between two points d (x,y) is defined in room R.

2. The patterns' or groups room A with elements $A_i$; the number of groups and the dimension of the room are finite or countable.

3. The stochastic function $\chi_x$, $x \in R$ with values in room $A_i$.

For every point x the distribution of $\chi_x$ is given in:

$$P(\chi_x = A_i) = p_i$$

$$\sum_{A_i \in A} p_i = 1$$

Exceptionally when $\chi_x$ is constant we get a determined $\varphi(R \to A)$

When R and A are known, and $\chi_x$ is undetermined we have the pattern recognition problem with the 'teacher'.

When R is known, and A is unknown then we have got a taxonomic problem.

When room R needs precision, we must first solve the problem of how to determine the informative subset of variables.

At the end a short histcry of the development of pattern recognition methods is presented.

# НЕПАРАМЕТРИЧЕСКИЙ МЕТОД САМООБУЧЕНИЯ В РАСПОЗНАВАНИИ ОБРАЗОВ

В.И. Васильев , В. В. Коноваленко

(Киев)

Большинство работ по распознаванию образов опирается на основополагающую гипотезу компактности, которая гласит: каждой совокупности объектов, образующих образ, соответствует компактное множество точек в пространстве изображений. Такая формулировка гипотезы компактности не полностью соответствует действительности, так как в ней не дается определение пространства изображений, в котором предполагается проявление свойства компактности.Очевидно,более точно эту гипотезу следует сформулировать так: если множество объектов представляет собой образ, то всегда найдется такое простанство, в котором этим объектам будет соответствовать компактное множество точек. Опираясь на эту формулировку, можно дать следующее определение абстрактного образа:абстрактный образ некоторого пространства - это множество объектов, не обязательно принадлежащих одному,определенному человеком, образу, которым в этом пространстве соответствует компактное множество точек. Тогда самообучение определим как процесс выделения абстрактных образов, в результате которого система приобретает только на основании смешанной обучающей выборки изображений способность к выработке одинаковых реакций на сходные (близкие) в заданном пространстве объекты.

Все известные алгоритмы самообучения способны выделять только абстрактные образы в заданных пространствах. Но это не снижает, а наоборот, иногда повышает ценность таких алгоритмов, так как часто задача именно в том и состоит,чтобы выделить в таком пространстве группы "похожих" объектов.

Примерами таких задач являются многие задачи социологии, технической и медицинской диагностики, в которых сам учитель не знает четкой классификации анализируемых объектов.

Предлагаемая группа алгоритмов отличается тем, что для решения задачи самообучения они не требуют задания извне числа классов (образов). Это является существенным преимуществом, так как в большинстве задач самообучения учителю неизвестно количество классов. Задачу самообучения сформулируем так.

Пусть имеется выборочное пространство $X$ наблюдений $\bar{x}_1$, $\bar{x}_2$, $\bar{x}_3$, ... над объектами $O_1$, $O_2$, ..., распределенными в соответствии с некоторой, неизвестной нам вероятностной мерой $P(X)$, такой что

$$P(X) = \int_X P(\bar{x})\,d\bar{x}. \qquad (1)$$

Пусть на $X$ задано множество $\mathcal{X}$ смешанной обучающей выборкой наблюдений $\bar{x}_1$, $\bar{x}_2$, ..., $\bar{x}_n$, распределенных в соответствии с мерой $P(X)$.

Требуется ввести на $X$ совокупность множеств

$$X = \bigcup_{i=1}^{M} M_i \; ; \; M_i \cap M_j = \emptyset \text{ при } i \neq j, \qquad (2)$$

где $M$ – заранее неизвестное число,
таких, чтобы каждое множество удовлетворяло определению абстрактного образа.

Одна из особенностей предлагаемого подхода к решению задачи самообучения в такой постановке состоит в том, чтобы отобразить выборочное пространство $X$ в некоторое другое, в котором бы свойство неоднородности значений меры близости между любой парой смешанной выборки наблюдений отобразилось в другое свойство, более удобное для целей распознавания. Если в качестве пространства отображения принять функцию плотности вероятности (фпв) смеси распределений, – а так как она неизвестна (1), – то ее оценку, вычисленную каким-либо способом для любой точки выборочного пространства, то, очевидно, выборочное пространство отобразится в множество значений оценки фпв смеси. При этом неравномерное распределение элементов $\bar{x} \subset \mathcal{X}$ отображается в многомодальное изменение оценки $P_n(\bar{x})$ фпв смеси $P(\bar{x})$. Причем каждая область "сгущений" элементов $\bar{x} \in \mathcal{X}$ (абстрактный образ) отображается в область значе-

ний $\hat{P}_n(\bar{x})$ , тяготеющих к одной и только одной ее моде, а области "разрежений" элементов $\bar{x} \in \bar{X}$ - границы между абстрактными образами - в "овраги" оценки фпв смеси.

Таким образом, если найдена оценка фпв смеси, то задачу самообучения в приведенной постановке можно считать решенной. Действительно, пусть на вход системы предъявляется описание объекта с неизвестной заранее классификацией. Этому описанию ставится в соответствие определенный элемент выборочного пространства, обозначим этот элемент $\bar{x}_0$ . Далее вычисляется значение оценки $\hat{P}_n(\bar{x}_0)$ , т.е. ищется отображение элемента $\bar{x}_0$ на множество значений оценки фпв смеси и из найденной точки на $\hat{P}_n(\bar{x})$ организуется поиск ближайшей моды этой функции. Ясно, что все отображения (а значит и соответствующие им точки выборочного пространства), "лежащие по одну сторону от дна оврага" $\hat{P}_n(\bar{x})$ , будут "скатываться" к одной и только одной моде $\hat{P}_n(\bar{x})$. Если считать реакцией системы указание номера моды оценки фпв смеси, к которой "тяготеет" отображение анализируемого объекта, то система будет одинаково реагировать на все множество отображений, расположенных с данной модой "по одну сторону от дна оврага". Так как "овраги" оценки фпв смеси являются отображением областей "разрежений" элементов выборочного пространства, то, очевидно, получаемые таким образом классы удовлетворяют определению абстрактного образа.

Очевидно, успех решения задачи самообучения при таком подходе зависит прежде всего от удовлетворительной оценки фпв смеси в любой точке выборочного пространства. В работе [1] предложена оценка, названная оценкой по методу нормальных вкладов, типа:

$$\hat{P}_n(x_0) = \frac{1}{n} \sum_{i=1}^{n} \hat{P}_n^*(x_0 - x_i) \qquad (3)$$

Смысл этой оценки состоит в том, что с каждым наблюдением смешанной обучающей выборки связана функция плотности нормального распределения $\hat{P}_n^*(x)$ , зависящая от $x_i$ , $i = 1,2,\ldots,n$ как от параметра, а значение оценки фпв смеси в любой точке $x_0$ выборочного пространства опреде-

ляется как среднее этих функций вклада. Там же исследуются статистические свойства оценки (3) и показано, что (3) является состоятельной оценкой $P_n(x_0)$ в любой точке $x_0 \in X$, непрерывной и не обладающей ложными экстремумами, если среднеквадратическое отклонение функции плотности нормального вклада вычисляется в соответствии с формулой:

$$\sigma_n = \frac{x_{(k)} - x_{(j)}}{2\sqrt{2}}, \qquad (4)$$

где $x_{(j)} < x_0 < x_{(k)}$, а $x_{(j)}$ и $x_{(k)}$ – порядковые статистики выборки.

В работе [2] предложена оценка фпв смеси, названная оценкой по методу ближайшего наблюдения, вида:

$$P_n(\bar{x}_0) = \frac{m \cdot \Gamma(m/2)}{2\pi^{m/2} \cdot r_n^m}, \qquad (5)$$

где $\Gamma(m/2)$ – гамма-функция,

$m$ – размерность выборочного пространства,

$r_n$ – радиус гиперсферы, равный среднему от суммы значений меры близости от точки $\bar{x}_0$, в которой вычисляется оценка $P_n(\bar{x}_0)$, до первого и второго ближайших к $\bar{x}_0$ наблюдений смешанной обучающей выборки.

Смысл этой оценки состоит в том, что с каждым наблюдением проверочной выборки связывается гиперсфера переменного радиуса $r_n$, значение которого зависит от распределения ближайших к данному наблюдению наблюдений смешанной обучающей выборки. Согласно (5), оценка фпв смеси оказывается обратно пропорциональной вычисленному указанным выше способом радиусу гиперсферы. Доказана состоятельность такой оценки в любой точке выборочного пространства. Там же приведена сравнительная оценка предложенной оценки (5) с некоторыми известными оценками фпв и показаны ее преимущества в смысле удобства практической реализации.

Обе оценки (3) и (5) фпв смеси являются непараметрическими, т.е. такими, которые не зависят от неизвестной нам функциональной формы смеси распределений, из которой извлечена смешанная обучающая выборка наблюдений.

Полное отсутствие информации о поведении оценки фпв смеси хотя бы в некоторой ограниченной ее области вокруг данной точки $\bar{x}_0 \in \mathcal{X}$, соответствующей наблюдению проверочной выборки, которое необходимо распознать, вынуждает ограничиться только регулярными методами поиска локальных экстремумов и только рабочими шагами перемещения в направлении к моде, равными пробным шагам. Это приводит к значительным потерям (например, машинного времени ) на поиск направления к ближайшей локальной моде оценки фпв смеси, которые возрастают с увеличением размерности выборочного пространства и объема обучающей выборки наблюдений. Поэтому с точки зрения упрощения практической реализации предложенного подхода к решению задачи самообучения имеет смысл попытаться ограничить удельный вес метода самообучения, основанного на использовании оценки фпв смеси для вычисления направления к ближайшей локальной моде и вывода отображения к этой моде.

Можно, например, попробовать ограничить применение метода самообучения только смешанной обучающей выборкой наблюдений. После того, как метод распределит наблюдения обучающей выборки по группам (абстрактным образам),после этого для распознавания каждого нового наблюдения можно применить другие методы распознавания, которые оказываются работоспособными на такой "обученной" обучающей выборке наблюдений.

Указанные выше непараметрические методы оценки фпв смеси применимы и к оценке условных фпв каждой из таких групп. Оценки априорных вероятностей каждой группы могут быть вычислены по данным объемов этих групп. Поэтому предлагаемый ниже метод распознавания наблюдений проверочной выборки по данным "обученной" обучающей выборки основан на применении байесового решающего правила:

$$\bar{x}_0 \in M_\alpha, \text{ если } P(M_\alpha)\ P(\bar{x}_0/M_\alpha) = \max_{j=1,\dots,M} P(M_j)\ P(\bar{x}_0/M_j) \qquad (6)$$

(при единичной матрице потерь).

Сформулируем задачу распознавания образов теперь так.

Пусть имеется выборочное пространство $\bar{X}$ наблюдений $\bar{x}_1$, $\bar{x}_2$, ... над объектами $O_1$, $O_2$, ..., распределенными в соответствии с некоторой, неизвестной нам вероятностной мерой $P(\bar{x})$.

Пусть на $\bar{X}$ задано множество $X$ обучающей выборкой независимых случайных наблюдений $\bar{x}_i = (x_{i1}, \ldots, x_{in})$, $i = 1,2, \ldots, n$, распределенных в соответствии с мерой $P(\bar{x})$. В обучающей выборке $n_j$ наблюдений принадлежат к группе $M_j$, $j = 1, \ldots, M$:

$$\sum_{j=1}^{M} n_j = n, \qquad M_i \cap M_j = \emptyset \qquad i \neq j. \tag{7}$$

Требуется вновь предъявленное наблюдение $\bar{x}$ отнести к одной из $M$ групп (абстрактных образов) таким образом, чтобы вероятность неправильной классификации (6) была минимальной.

Решение задачи распознавания в такой постановке сводится к нахождению оценок априорных вероятностей $P(M_j)$, $j = 1, \ldots, M$ групп, оценок условных фпв $P(\bar{x}/M_j)$ по обучающей выборке наблюдений и использовании этих оценок в решающем правиле (6).

Оценку априорной вероятности будем вычислять по формуле:

$$\hat{P}(M_j) = \frac{n_j}{n}, \qquad j = 1, \ldots, M. \tag{8}$$

Методика вычисления оценок условных фпв имеет много общего с методикой вычисления оценки фпв смеси. Поэтому при реализации распознающей программы, например, на ЭЦВМ, удобно применять на этапе самообучения (когда смешанная обучающая выборка наблюдений расчленяется на группы-абстрактные образы) и на этапе обучения (когда решается задача отнесения наблюдения проверочной выборки к тому или иному абстрактному образу из заданной их совокупности) один и тот же метод непараметрической оценки фпв. В результате, усложнение программы вычислений получается незначительным, а выигрыш в потерях машинного времени на классификацию наблюдений проверочной выборки - существенным.

Рассмотрим оценки условных фпв по методу ближайшего наблюдения [2] и проведем исследование их статистических свойств.

Конструирование предлагаемой ниже оценки условной фпв $P_n(\bar{x}_0/M_j)$, $j = 1,\dots, M$ осуществляется в следующей последовательности.

I. Определение значения меры близости к первому ближайшему к данной точке $\bar{x}_0$, в которой требуется вычислить $M$ оценок условных фпв, наблюдения $\bar{x}_i$ обучающей выборки, принадлежащего группе $n_j$, $j = 1,\dots, M$:

$$d(\bar{x}_0, \bar{x}_i) = \min_{i=1,\dots,n_j} d(\bar{x}_0, \bar{x}_i), \quad j = 1, \dots, M. \tag{9}$$

2. Определение значения меры близости ко второму ближайшему к данной точке $\bar{x}_0$ наблюдению обучающей выборки из той же группы:

$$d(\bar{x}_0, \bar{x}_t) = \min_{t=1,\dots,i-1,i+1,\dots,n_j} d(\bar{x}_0, \bar{x}_t), \quad j = 1, \dots, M \tag{10}$$

3. Вычисление величины

$$r_{n_j} = \frac{1}{2}\left[d(\bar{x}_0, \bar{x}_i) - d(\bar{x}_0, \bar{x}_t)\right]. \tag{11}$$

4. Установление гиперсферы радиуса $r_{n_j}$ с центром в точке $\bar{x}_0$:

$$S_{r_{n_j}, \bar{x}_0} = \{\bar{x} : d(\bar{x}_0, \bar{x}) \le r_{n_j}\}, \quad j = 1, \dots, M. \tag{12}$$

5. Вычисление объема гиперсферы $V_{r_{n_j}, \bar{x}_0}$ по формуле [3]:

$$V_{r_{n_j}, \bar{x}_0} = \frac{2\pi^{m/2} r_{n_j}^{m}}{m \cdot \Gamma(m/2)}, \quad j = 1, \dots, M. \tag{13}$$

6. Вычисление оценки условной фпв как отношения вида

$$P_n(\bar{x}_0/M_j) = \frac{P(S_{r_{n_j}, \bar{x}_0})}{V_{r_{n_j}, \bar{x}_0}} \tag{14}$$

где $P(S_{r_j, \bar{x}_0})$ – доля вероятностной меры $P(\bar{X})$, попадающей в гиперсферу $S_{r_j, \bar{x}_0}$.

Покажем, что отношение (14) действительно является оценкой условной фпв $P(\bar{x}_0/M_j)$, $j = 1, \ldots, M$ и что эта оценка состоятельна, т.е. докажем, что выполняется равенство

$$\lim_{n \to \infty} P\left[\frac{P(S_{r_j, \bar{x}})}{V_{r_j, \bar{x}_0}} = P(\bar{x}_0/M_j)\right] = 1. \qquad (15)$$

Это доказательство основано на применении теории покрытий или непараметрических областей [4]. Эта теория основана на порядковых статистиках, которые всегда можно получить по данным обучающей выборки наблюдений. В $m$-мерном случае порядковые статистики строятся при помощи упорядочивающих функций. Не накладывается жестких ограничений на вид упорядочивающей функции и на их количество. Примем в качестве упорядочивающих функций $M$ функций вида (11). Тогда эти $M$ функций определят в $m$-мерном выборочном пространстве $M$ гиперсфер с центром в точке $\bar{x}_0$ и различными радиусами $r_{x_j}$, $j = 1, \ldots, M$. Каждая такая гиперсфера (выборочный блок) содержит одно, ближайшее к точке $\bar{x}_0$ наблюдение обучающей выборки из соответствующей группы. Очевидно, упорядоченный ряд чисел $r_{x_j}$, $j = 1, \ldots, M$ задает упорядоченный по отношению к точке $\bar{x}_0$ ряд $m$-мерных наблюдений обучающей выборки, в чем, собственно и состоит назначение упорядочивающих функций.

С каждым из полученных таким образом выборочным блоком связана случайная переменная, называемая долей выборочного блока. Согласно теории покрытий распределение долей не зависит от вида условных фпв и всегда подчинены бета-распределению. При вычислении оценки условной фпв $P_n(\bar{x}_0/M_1)$, рассматривается выборочный блок $S_{r_{x_1}, \bar{x}_0}$, который определяет значение случайной переменной – доли $P(S_{r_{x_1}, \bar{x}_0})$ и объем в выражении (14). При определении $P_n(\bar{x}_0/M_2)$ рассматривается доля $P(S_{r_{x_2}, \bar{x}_0})$ и объем выборочного блока $S_{r_{x_2}, \bar{x}_0}$ и т.д. Следовательно, при вычислении любой из $M$ оценок

условных фпв используется только один из $M$ выборочных блоков, который каждый раз оказывается первым по отношению к точке $\bar{x}_0$. Тогда, на основании следствия, которое легко можно получить из одной теоремы теории покрытий, приходим к выводу, что каждый выборочный блок ( гиперсфера) $S_{n_j,\bar{x}_0}$ подчинен бета-распределению с математическим ожиданием

$$M[P(S_{n_j,\bar{x}_0})] = \frac{1}{n_j+1}, \quad j=1,\dots,M \tag{16}$$

и дисперсией

$$D[P(S_{n_j,\bar{x}_0})] = \frac{n_j}{(n_j+1)^2(n_j+2)}, \quad j=1,\dots,M. \tag{17}$$

Таким образом, оценки условных фпв $P_n(\bar{x}/w_j), j=1,\dots,M$ всегда подчинены бета-распределению с параметрами (16) и (17), независимо от положения точки $\bar{x}_0$ в выборочном пространстве $X$ и функционального вида неизвестных условных фпв. Поэтому такие оценки типа (14) называются непараметрическими. Естественно, доказательство состоятельности оценок (14) основано на использовании свойств доли $P(S_{n_j,\bar{x}_0})$.

Так как дисперсия (17) доли $P(S_{n_j,\bar{x}_0})$ асимптотически стремится к нулю, то применение закона больших чисел позволяет записать:

$$\lim_{n\to\infty} P[P(S_{n_j,\bar{x}_0}) < \varepsilon] = 1. \tag{18}$$

Из этого равенства следует, что при $n\to\infty$ доля вероятностной меры $P(S_{n_j,\bar{x}_0})$, попадающая в гиперсферу $S_{n_j,\bar{x}_0}$, уменьшается до нуля. Это случается тогда, если при $n\to\infty$ происходит "стягивание" гиперсферы $S_{n_j,\bar{x}_0}$ в точку. Однако, "стягивание" гиперсферы может произойти только в том случае, если при $n\to\infty$ уменьшается ее объем $V_{n_j,\bar{x}_0}$. Объем же уменьшится только тогда, если при $n\to\infty$ радиус гиперсферы $r_j$, $j=1,\dots,M$ стремится к нулю, что имеет место, если взять предел от выражения (12).

Таким образом, точка, к которой "стягивается" доля $P(S_{n_j,\bar{x}_0})$ и объем $V_{n_j,\bar{x}_0}$ гиперсферы $S_{n_j,\bar{x}_0}$, $j = \overline{1,}$ ..., $M$ при $n \to \infty$ это точка $\bar{x}_0$. Значит отношение

$$\lim_{n\to\infty} \frac{P(S_{n_j,\bar{x}_0})}{V_{n_j,\bar{x}_0}} = p(\bar{x}_0/M_j), \quad j = \overline{1,..,M} \tag{19}$$

представляет собой функцию плотности в точке $\bar{x}_0$ $m$-мерного выборочного пространства. Этого достаточно, чтобы заключить, что равенство (15) справедливо.

Итак, выражение (15) представляет собой состоятельную оценку условных ФПВ в любой точке выборочного пространства. Однако, воспользоваться этой формулой невозможно, так как доля вероятностной меры $P(S_{n_j,\bar{x}_0})$, входящая в нее, так же неизвестна, как и сама вероятностная мера $P(\bar{X})$. Необходимо найти оценку $\hat{P}(S_{n_j,\bar{x}_0})$ и заменить ею долю $P(S_{n_j,\bar{x}_0})$ в (14), чтобы полученная таким образом новая оценка

$$p_n(\bar{x}_0/M_j) = \frac{\hat{P}(S_{n_j,\bar{x}_0})}{V_{n_j,\bar{x}_0}}, \quad j = \overline{1,..,M} \tag{20}$$

являлась состоятельной оценкой $P(\bar{x}_0/M_j)$, $j = \overline{1,...,M}$:

$$\lim P\left[\frac{P(S_{n_j,\bar{x}_0})}{V_{n_j,\bar{x}_0}} = P(\bar{x}_0/M_j)\right] = 1. \tag{21}$$

Примем в качестве оценки $\hat{P}(S_{n_j,\bar{x}_0})$ значение, которое принимает выборочная функция распределения для данного выборочного блока $S_{n_j,\bar{x}_0}$, $j = \overline{1,...,M}$:

$$\hat{P}(S_{n_j,\bar{x}_0}) = \frac{1}{n_j}, \quad j = \overline{1,..,M}. \tag{22}$$

Для доказательства (21) достаточно указать, что в соответствии с законом больших чисел выборочная функция распределения асимптотически сходится к теоретической функции распределения, т.е. имеет место сходимость по вероятности оценки $\hat{P}(S_{n_j,\bar{x}_0})$ к доле $P(S_{n_j,\bar{x}_0})$, $j = \overline{1,...,M}$.

Используя в (6) вместо априорных вероятностей $P(M_j)$ и условных фпв $p(\bar{x}_0/M_j)$, $j = \overline{1,...,M}$ их оценки (8) и (20), получим относительно любого наблюдения проверочной выборки решающее правило:

$$\bar{x}_0 \in M_k \quad \text{если} \quad \hat{P}(M_k)\cdot P_n\left(\bar{x}_0/M_k\right) = \max_{j=1,\dots,m} \hat{P}(M_j)\cdot P_n\left(\bar{x}_0/M_j\right) \qquad (23)$$

что эквивалентно, с учетом (I3), (22)

$$\bar{x}_0 \in M_k \text{ если } \frac{1}{r_{M_k}} = \max_{j=1,\dots,m} \frac{1}{r_{M_j}} \qquad (24)$$

Решающее правило (24) инвариантно к априорным вероятностям групп (абстрактных образов). Для некоторых типов задач это свойство может оказаться положительным. Например, пусть $P(M_1) = P(M_2) = \frac{1}{2}$ а объемы обучающих выборок для каждой группы равны $n_1 = k$ и $n_2 = 3k$. В этом случае оценки априорных вероятностей групп равны $\hat{P}(M_1) = \frac{1}{4}$ и $\hat{P}(M_2) = \frac{3}{4}$ и применение байесового решающего правила (23) приводит к значительным ошибкам. В то же время решающее правило (24) может дать значительно лучшие результаты.

Практически решающее правило (24) указывает, что любое наблюдение проверочной выборки следует относить к той группе, одно из наблюдений обучающей выборки которой расположено ближе всего к нему. В этом смысле предложенное решающее правило совпадает с известным ранее методом распознавания Ковера Т.М. и Харта П.Е. [5].

Предложенный подход к решению задачи самообучения был применен к решению задачи диагностики заболеваний "Инфаркт миокарда".

Смешанная обучающая выборка включала 98 больных, проверочная выборка - I92 больных заболеванием "Инфаркт миокарда".

Состояние каждого больного описывалось 240-мерным вектором ( в машинных символах). Все больные были распределены врачом по I4 диагнозам (эта информация не использовалась при решении задачи обучения). Каждый диагноз врача был представлен таким количеством больных:

Таблица I

| код диагноза врача | количество больных | код диагноза врача | количество больных |
|---|---|---|---|
| 02 | 2 | II | 2 |
| 03 | I06 | I2 | I |
| 04 | 2I | I4 | I |
| 05 | 29 | I5 | I |
| 06 | I8 | I6 | 2 |
| 07 | 23 | 20 | I |
| I0 | 79 | 23 | 4 |

Как следует из приведенной таблицы I, состав диагнозов крайне неоднородный. В обучающую выборку путем случайного отбора были выделены больные первых семи диагнозов, в количествах, пропорциональных их составу.

Программа, реализующая самообучение по методу нормальных вкладов, распределила больных проверочной выборки по заданным их описаниям на I9 групп следующим образом:

Таблица 2

| № пп | | | | |
|---|---|---|---|---|
| № группы | Количество больных в группе | Количество больных с одинаковым диагнозом врача | | код диагноза врача |
| I | 27 | 22 | 0,82 | I0 |
| 2 | I3 | 8 | 0,62 | I0 |
| 3 | 6 | 3 | 0,50 | I0 |
| 4 | I4 | II | 0,79 | I0 |
| 5 | I3 | 5 | 0,37 | I0 |
| 6 | I0 | 7 | 0,70 | 07 |
| 7 | 6 | 4 | 0,67 | 07 |
| 8 | 22 | I7 | 0,77 | 03 |
| 9 | 9 | 7 | 0,78 | 03 |
| I0 | 9 | 7 | 0,78 | 03 |
| II | 9 | 9 | I,00 | 03 |
| I2 | 8 | 5 | 0,63 | 03 |
| I3 | 6 | 5 | 0,83 | 03 |
| I4 | 23 | I6 | 0,70 | 03 |
| I5 | 7 | 5 | 0,7I | 03 |
| I6 | 4 | 2 | 0,50 | 04 |
| I7 | 5 | 3 | 0,60 | 04 |
| I8 | 6 | 6 | I,00 | 05 |
| I9 | 6 | 4 | 0,67 | 06 |

В графе 4 данной таблицы приведено отношение числа больных с одинаковым диагнозом врача к общему количеству больных в каждой группе. В графе 5 - код диагноза, который поставил врач таким больным в каждой группе. Как следует из таблицы 2, средняя точность распознавания, оцениваемая числом совпавших диагнозов машины и врача ( по отношению к общему количеству больных проверочной выборки) равна 75% . Специалисты-медики положительно оценили такие результаты машинной диагностики, так как даже высококвалифицированный врач часто оказывается не в состоянии достичь такого процента правильных диагнозов в заданном пространстве. Пространство описаний болезней "Инфаркт миокарда" было задано таким, что оно не содержало данных электрокардиограмм и других инструментальных методов анализа и постановки диагноза.

Эсперимент проводился на ЭЦВМ "БЭСМ-6". Программа составлена на языке "БЭСМ-АВТОКОД". Время постановки машиной диагноза для одного больного составило от I0 сек. до 3-х минут. Общее время постановки диагнозов выборке из 200 больных составило 5 часов машинного времени.

Литература

I. Васильев В.И., Коноваленко В.В. Самообучение в задаче распознавания образов, сб."Техническая кибернетика", Институт Кибернетики АН УССР, вып. 2, I970 г.

2. Коноваленко В.В. Непараметрическая оценка функции плотности вероятности по методу ближайшего наблюдения, сб. "Техническая кибернетика", Институт Кибернетики АН УССР, вып. , I97I г.

3. Корн Г., Корн Т., Справочник по математике , "Мир", I970г.

4. Уилкс С. Математическая статистика, "Мир", I967 г.

5. Cover T.M., Hart P.E., Nearest neighbour pattern classification, IEEE Trans. Inform. Theory, 1967, 13,N 1 ( реферат в ж. "Техническая кибернетика", № 27, I967 г.)

# MITTEPARAMEETRILINE ISEÕPPIMISE MEETOD KUJUNDITE ERISTAMISES

V.I. Vassiljev , V.V, Konovalenko
(Kiiev)

Resümee

Artiklis defineeritakse abstraktne kujund tavalisest käsitlusest mõnevõrra üldisemalt - see on objektide hulk, mis ei tarvitse moodustada antud ruumis kompaktset hulka, kuid alati leidub mingi ruum, milles see punktihulk osutub teatud mõttes kompaktseks.

Iseõppimine on siis abstraktsete kujundite eraldamise protsess,mille käigus antud õpperühma põhjal süsteem reageerib selles ruumis lähedastele objektidele ühesugusel viisil.

Esitatakse rühm iseõppimise algoritme, mis ei nõua kujundite ega nende arvu etteandmist. Selleks kujutatakse väljavõtete ruum uude ruumi, milles defineeritud läheduse mõõt on paremini kooskõlas antud kujundite eristamise ülesandega. Ühe võimalusena on vaadeldud jaotuste segu tihedusfunktsiooni kujutisruumi mõõduna; probleem taandub siis tihedusfunktsiooni statistilisele hindamisele; selleks esitatakse mitteparameetriline meetod,mis tugineb järkstatistikutele.

Lõpuks tuuakse rakenduslik näide, kus toodud metoodikat kasutatakse südamehaigete diagnoosimiseks 240 tunnuse põhjal. Õpperühm sisaldas 98,kontrollrühm 192 haiget(müokardi infarkt). Arsti ning masina diagnooside ühtelangevus moodustas keskmiselt 75%.

## NON-PARAMETRIC SELF-LEARNING METHOD IN PATTERN RECOGNITION

V.I. Vassilyev, V.V. Konovalenko
(Kiev)

Summary

In the article the abstract pattern is defined in a more general sense than usually it is a set of objects that needs not form a compact set in a given room although there always exists a room where the set of points will be compact.

Self-learning is thus a process of recognizing abstract patterns in the course of which the system responds to similar objects in agiten room in a similar way.

A group of self-learning algorithms is distinguished which do not demand any prior knowledge of patterns or their number. For that purpose a sample room is projected into another room where the defined measure of proximity is in a better accordance with a given task of pattern recognition. Such measure may be a mixture of density functions. The problem may thus be reduced to the statistical estimation of the density function; this estimation is based on the non-parametric method.

At the end an example of the practical use of the method is given. The method is applied for diagnosing heart diseases. The subject group consisted of 98 patients, the reference group of 192 patients. The diagnoses of a physician and of a machine coincided in 75 per cent of cases.

# УЧЕТ ПРЕДЫСТОРИИ ПРИ РАСПОЗНАВАНИИ ОБРАЗОВ

В.И.Васильев, В.Е. Реуцкий

(Киев)

## I. Вступление

На практике часто встречаются задачи распознавания объектов в процессе их движения или последовательного развития. При этом объекты переходят из одного состояния в другое, последовательность которых определяет принадлежность объекта к тому или иному классу. Каждое состояние объекта определяется совокупностью признаков. В своей эволюции вектор, соответствующий описанию объекта, вычерчивает определенную траекторию в пространстве признаков. В таком случае задачу распознавания можно определить как классификацию совокупностей состояний или как классификацию траекторий. Например, тело движущееся в пространстве, последовательно наблюдается в трех положениях (ракурсах) А, Б, В. В первом случае необходимо распознать объект по совокупности всех ракурсов, т.е. их наличие в описании тела, а во втором случае необходимо учитывать не только наличие ракурсов, но и порядок их смены. Такая постановка задачи предусматривает, что на результат распознавания в большей мере влияет история эволюции объекта (предыстория), чем какое-либо отдельное состояние или ракурс. Здесь сразу же возникает много неясных вопросов и в частности вопрос об обпределении момента , после которого дальнейшее наблюдение за поведением объекта для его классификации нецелесообразно.

## 2. Учет совокупности состояний

Пусть все пространство признаков $\{X\}$ ($i = 1, 2, \ldots n$) разбито на $M$ областей так, что каждой области соответствует какое-то характерное состояние объекта и каждая область характеризуется своим описанием $X_\nu = x_{i\nu}, x_{2\nu}, \ldots x_{n\nu}$;

$j = 1, 2, \ldots N$ . В частности области могут характеризоваться полюсами, а их границы определяться силами притяжения этих полюсов. Будем считать, что объект в процессе наблюдения изменяется так, что соответствующая представляющая точка в пространстве признаков в любой момент времени обязательно находится в одной из областей $M$ . Другими словами, факт принадлежности представляющей точки к одной из областей является одним из множества единственно возможных и несовместимых событий $X$ . Требуется на основании совокупности состояний, пройденных точкой за время наблюдения, определить к какому из $\{V_j\}$ ( $j = 1,2, \ldots m$ ) классов принадлежит объект.

Пусть совокупность результатов наблюдений $X$ за объектом $V$ представляет собой выборку объемом $N$ . В общем случае

$$X = (X_1, X_2, \ldots, X_N). \qquad (1)$$

Каждое наблюдение состоит из набора признаков и отличается от других наблюдений комбинацией этих признаков. Момент прекращения наблюдения за объектом можно определить исходя из того, что совокупность наблюдений в своей последовательности достаточна для вынесения определенного решения. Поставленная задача очень близка к известной задаче последовательного анализа (1). Действительно, если задачу распознавания рассматривать с точки зрения математической статистики, то она полностью совпадает с задачей различения гипотез. Каждому классу ставится в соответствие определенная мера, заданная на пространстве описаний (признаков),а параметром,по которому отличаются эти меры, является номер класса. На основании наблюдений следует предпочесть одну из гипотез о значении неизвестного параметра, т.е. о принадлежности объекта к одному из классов.

В классической постановке задача различения гипотез предполагает распределения вероятностей в пространстве описаний известными. В задаче распознавания эти распределения определяются в процессе обучения. Будем считать, что в процессе обучения определены значения $P(x_i/V_j)$ и $P(X_v/V_j)$

для всех значений $i, j, v$ , ( $j = 1, 2, \ldots m$ ). Здесь $P(x_i / V_j)$ – плотность вероятности (вероятность) появления признака $x_i$ при условии класса $V_j$ , а $P(X_v / V_j)$ – плотность вероятности (вероятность) появления наблюдения $X_v$ при условии класса $V_j$ . Причем выражения вида $P(x/V)$ следует понимать либо как плотность вероятности, либо как вероятность соответствующего описания, когда множества $x$ и $X$ дискретны. Будем также считать, что совокупность наблюдений $x_i$ представляют собой множество независимых испытаний, а признаки $x$ независимы. В силу независимости наблюдений вероятность получения выборки, совпадающей с наблюдаемой, определяется как

$$P(X_1, X_2, \ldots, X_N) = P(X_1)\,P(X_2)\ldots P(X_N), \qquad (2)$$

где

$$P(X_v) = P(x_{1v})\,P(x_{2v})\ldots P(x_{nv}) = P(x_{iv}). \qquad (3)$$

Решение о принятии какой-либо гипотезы выносится на основании сравнения вероятностей вида

$$P(X/V_j) = P(X_1/V_j)\,P(X_2/V_j)\ldots P(X_N/V_j) = P(X/V), \qquad (4)$$

где

$$P(X/V_j) = P(x_{1v}/V_j)\,P(x_{2v}/V_j)\ldots P(x_{nv}/V_j). \qquad (5)$$

Рассмотрим случай, когда все объекты необходимо разделить на $M$ классов. Для этого необходимо сформулировать план последовательной проверки выбора из $m$ взаимно исключающих и исчерпывающих все возможные случаи гипотез $V_1$ , $V_2$ , ... , $V_m$ . В общих чертах план может быть описан следующим образом. Формулируется правило принятия одного из $m + 1$ решений на каждом $N$ -этапе наблюдений:

1. Закончить наблюдение принятием гипотезы $V_1$ .
2. Закончить наблюдение принятием гипотезы $V_2$ .

. . . . . .

$m$ Закончить наблюдение принятием гипотезы $V_m$.

$m+1$ Продолжать наблюдения.

План последовательной проверки будем считать оптимальным, если он обеспечивает заданную вероятность ошибки при наименьшем числе наблюдений. В дальнейшем оптимальность будет пониматься в этом смысле без специальных оговорок. Для построения оптимального последовательного критерия обратимся к следующей лемме [1].

Лемма. Пусть $X_1, X_2, \ldots$ последовательность переменных, а $P_{1N}$ ($N = 1, 2, \ldots$) совместная плотность вероятности $X_1, X_2, \ldots, X_N$ при условии, что верна гипотеза $V_1$ и $P_{2N}$ – плотность вероятности при условии, что верна гипотеза $V_2$. Тогда при условии, что верна гипотеза $V_2$, вероятность того, что неравенство

$$\frac{P_{1N}(X_1, X_2, \ldots, X_N)}{P_{2N}(X_1, X_2, \ldots, X_N)} < A; \quad \text{где} \quad 1 < A = const; \tag{6}$$

будет выполнено для всех значений $N$, больше или равна $1 - \frac{1}{A}$.

Сущность леммы состоит в том, что вероятность принятия гипотезы $V_2$ (вероятность нарушения неравенства) при условии справедливости $V_1$ тем меньше, чем больше значение A и наоборот, если A близко к единице, вероятность принятия гипотезы $V_2$, когда верна $V_1$, наиболее велика, т.е. мала вероятность выполнения неравенства.

Если принимать решение только в том случае, когда $\frac{P_{1N}}{P_{2N}} \geq A$, причем это решение всегда будет в пользу гипотезы $V_1$, а решение $V_2$ вообще не выносится (ошибка второго рода запрещена), то вероятность неправильного решения не будет превышать $\frac{1}{A}$. Если принять $A = \frac{1}{P_0}$ (где $P_0$ – допустимая вероятность ошибки), то вероятность неправильного решения не превысит $P_0$. Лемма справедлива для двух простых гипотез и дает рекомендации относительно выбора порога A.

На основании этой леммы можно построить план последовательной проверки для выбора одной из $m$ взаимоисключающих гипотез. Будем считать, что вероятности $P(x_i / V_j)$

определены в период обучения для всех значений $i = 1, 2, \ldots n$ и $j = 1, 2, \ldots m$, причем вероятности $P(X_i/V_j)$ определяются как в (4). Пусть

$$P(X/V) = P(X_1, X_2, \ldots, X_N/V) = P(X_1/V)\,P(X_2/V)\ldots P(X_N/V) \qquad (7)$$

есть вероятность того, что объект с описанием $X$ соответствует классу $V$. Сформулируем три последовательных критерия, которые непосредственно вытекают из вышеприведенной леммы.

Критерий 1. На каждой стадии наблюдений $N$ вычисляются $m$ отношений вероятностей

$$\frac{P_N(X/V_j)}{P_N(X/\bar{V}_j)}, \qquad (8)$$

где $\bar{V}_j = V^* - V_j = \bigcup_{\substack{i=1 \\ i \neq j}}^{m} V_i$ – дополнение к гипотезе $V_j$,

$V^* = \bigcup_{i=1}^{m} V_i$ – полная группа несовместимых событий,

$P_N(X/\bar{V}_j)$ – вероятность $X$ при условии любой из гипотез кроме $V_j$.

Если

$$\frac{P_N(X/V_j)}{P_N(X/\bar{V}_j)} < A \qquad (9)$$

для всех $j = 1, 2, \ldots m$, то наблюдение продолжается до тех пор, пока для какого-либо значения $j = k$ будет выполнено неравенство

$$\frac{P_N(X/V_k)}{P_N(X/\bar{V}_k)} \geq A \qquad (10)$$

и в этом случае предпочтение отдается гипотезе $V_k$. Если на одном и том же шаге неравенство (10) выполняется сразу для нескольких значений $j$, то преимущество отдается той гипотезе, для которой оно максимально. Назовем этот критерий последовательным критерием отношения вероятностей гипотез к их дополнениям или сокращенно, критерием простого дополнения.

Критерий 2. На каждом шаге наблюдения вычисляются отношения вероятностей

$$\frac{\max_j P_N(X/V_j)}{P_N(X/\bar{V}_j)}. \qquad (11)$$

где $V_j$ – гипотеза, при которой $P(X/V_j)$ достигает максимума, а $\bar{V}_j$ – дополнение к гипотезе $V_j$. Наблюдения продолжаются, если

$$\frac{\max_j P_N(X/V_k)}{P_N(X/\bar{V}_k)} < A. \qquad (12)$$

Если же для $j = k$

$$\frac{\max_j P_N(X/V_k)}{P_N(X/\bar{V}_k)} \geq A, \qquad (13)$$

то процесс наблюдения оканчивается принятием гипотезы $V_k$. Этот критерий будем называть критерием отношения наиболее правдоподобной гипотезы к своему дополнению, или коротко, критерием минимального дополнения.

Критерий 3. На каждой стадии наблюдения $N$ вычисляется отношение вероятностей

$$\frac{\max_j P_N(X/V_j)}{\max_{j\neq k} P_N(X/V_j)}, \qquad (14)$$

где $k$ – индекс, при котором достигается максимум $P(X/V)$. Если же

$$\frac{\max_j P_N(X/V_j)}{\max_{j\neq k} P_N(X/V_j)} < A, \qquad (15)$$

то эксперимент продолжается и производится дополнительное $m+1$ наблюдение. Если же

$$\frac{\max_j P_N(X/V_j)}{\max_{j\neq k} P_N(X/V_j)} \geq A \qquad (16)$$

то процесс наблюдения оканчивается принятием гипотезы $V_k$. Этот критерий назовем критерием наиболее правдоподобных гипотез.

Следует отметить, что последние два критерия дают требуемый результат только при равновероятных гипотезах. В (3) доказаны следующие теоремы.

Теорема 1. Последовательный критерий простого дополнения является оптимальным в смысле минимума среднего числа наблюдений при заданной взвешенной средней вероятности ошибки.

Теорема 2. Если конкурирующие гипотезы равновероятны, то последовательный критерий минимального дополнения может обеспечить заданный уровень вероятности ошибки и в этом случае он будет оптимальным.

Теорема 3. В случае равновероятных гипотез последовательный критерий наиболее правдоподобных гипотез требует большего числа наблюдений, чем критерий минимального дополнения, но при одинаковых значениях порога А этот критерий обладает большей мощностью.

Теорема 4. В случае равновероятных гипотез критерий наиболее правдоподобных гипотез является оптимальным в смысле минимума среднего числа наблюдений и обеспечивает среднее значение

$$\beta = \min\left[\max \beta_i (V_i \in \bar{V}_j)\right]. \qquad (17)$$

Так как наибольший интерес практически представляет критерий наиболее правдоподобных гипотез, то здесь приведено доказательство только четвертой теоремы, где $\beta_i$ – вероятность ошибки, состоящей в том, что в случае принятия гипотезы $V_j$, оказывается справедливой гипотеза $V_i \in \bar{V}_j$ ($\bar{V}_j$ – дополнение к гипотезе $V_j$).

Доказательство теоремы 4. Будем считать, что все конкурирующие гипотезы составляют множество $\bar{V}^* = \bigcup_{i=1}^{m} V_i$. Задача состоит в том, чтобы проверить одну из простых гипотез $V_j$ относительно множества конкурирующих $\bar{V}_j = \bigcup_{i=1}^{m} V_i$. Для простоты положим $i = 1$ (т.е. $V_j = V_1$), а $\bar{V}_j = \bar{V}_1 = \bigcup_{i=2}^{m} V_i$.

Класс последовательных критериев, обеспечивающих заданную взвешенную среднюю вероятность ошибки, включает в себя, по крайней мере, столько критериев, сколько может быть весовых функций $F(V)$, удовлетворяющих условию

$$\int_{\bar{V}_1} F(V)dV = 1; \quad \sum_{i=2}^{m} F(V_i) = 1. \qquad (18)$$

Каждый из критериев однозначно определяется весовой функцией $F(V)$ и выбором порога А. Необходимо выбрать весовую функцию так, чтобы при заданном пороге А, максималь-

ная величина $\beta(V)$ по всей области $W$ (область преобладания гипотезы $V_2$, определяемая неравенством $p_{1N}/p_{2N} \leq A$) была бы минимальной.

Предположим, что имеется критерий, основанный на сравнении заданной постоянной A с отношением

$$\frac{p_{1N}}{p_{2N}} = \frac{p(X/V_i)}{\int_{V_2} p(X/V)F^*(V)dV} \approx \frac{p(X/V_i)}{\sum_{i=2}^{m} p(X/V_i)F^*(V_i)}, \qquad (19)$$

причем весовая функция $F^*(V)$ такая, что

$$\int_W F^*(V)dV = \int_{V_2} F(V)dV \approx \sum_{V_i \in W} F^*(V_i) = \sum_{i=2}^{m} F^*(V_i) = 1; \qquad (20)$$

и

$$\int_{V_2} \beta^*(V)F^*(V)dV \approx \sum_{i=2}^{m} \beta^*(V_i)F^*(V_i) = \bar{\beta}^*; \qquad (21)$$

где $V_2$ – область, включающая в себя все точки, лежащие вне области принятия гипотезы $V_1$, определяемой заданным критерием, а $W$ множество точек, для которых

$$\beta^*(V) = \max_{V_i \in V_2} \beta(V_i) = const. \qquad (22)$$

Такая весовая функция представляет собой предел функции, определяемой (I8) при условии, что она принимает значение 0 в любой точке области $V_2$, не входящей в $W$. Так как при выборе функции, определяемой условия (I8) не нарушаются, рассматриваемый критерий должен гарантировать предел средней взвешенной ошибки, равной $\bar{\beta}^*$. Так как по условию $\beta^*(V)$ в области $W$ постоянна и максимальна, (22) можно записать

$$\bar{\beta}^* = \max_{V_i \in V_2} \beta(V_i) = \beta^*(V) = \int_{V_2} \beta(V)F^*(V)dV. \qquad (23)$$

Теперь рассмотрим этот же критерий при той же постоянной A, но введем любую другую весовую функцию $F^{**}(V)$, подчиненную условию (I8). В (I) показано, что выбор весовой функции $F(V)$ влияет только на вид функции $\beta(V)$, но

в рамках заданного критерия и заданной постоянной $A$ никак не влияет на величину средней взвешенной ошибки. Отсюда следует, что

$$\bar{\beta}^{*} = \bar{\beta}^{**} = \int_{V_2} \beta^{*}(V) F^{*}(V) dV = \int_{V_2} \beta^{**}(V) F^{**}(V) dV, \tag{24}$$

но

$$\max_{V_i \in V_2} \beta^{**}(V_i) \geq \bar{\beta}^{**}, \tag{25}$$

а по условию (22)

$$\bar{\beta}^{*} = \max_{V_i \in V_2} \beta^{*}(V_i). \tag{26}$$

Сравнивая (22), (24), (25) можно заметить, что

$$\max_{V_i \in V_2} \beta^{**}(V_i) \geq \max_{V_i \in V_2} \beta^{*}(V_i). \tag{27}$$

На весовую функцию $F^{**}(V)$ пока не накладывалось никаких ограничений кроме (I8). Значит неравенство (27) справедливо для всех

$$F^{**}(V) \neq F(V). \tag{28}$$

Таким образом, если критерий (I9) содержит весовую функцию, выбранную в соответствии с (20 - 22), то такой критерий обеспечивает

$$\min_{j} [\max_{i} \beta_{ij} (V \in \bar{V}_j)]. \tag{29}$$

Но так как этот критерий относится к классу последовательных критериев отношения вероятностей, то он требует в среднем наименьшего числа наблюдений среди всех возможных критериев (I).

Для фиксированного значения $X$ максимальная вероятность ошибки $\beta_{ij}$ будет соответствовать тем значениям $i$, для которых

$$P(X/V \in V_j) = \max_{i}, \tag{30}$$

а это значит, что весовая функция $F^{*}(V_i)$ должна обращаться в нуль повсюду, кроме тех $i$, для которых выполняется равенство (30).

Для $i$, удовлетворяющих равенству (30), эта функция равна единице. При этом в критерии (I9) знаменатель принимает вид

$$\sum_{i=1}^{m} P(X/V_i)F^*(V_i) = \sum_{i=1}^{m} \max_i P(X/V_i). \qquad (31)$$

Практически равенство (31) выполняется только для одного значения $i$ и в этом случае критерий (19) ничем не отличается от критерия наиболее правдоподобных гипотез, а так как критерий (19) обеспечивает

$$\min[\max_i \rho_i(V_i \in \bar{V}_j)]$$

при наименьшем среднем числе наблюдений, то теорема доказана.

### 3. Учет последовательности пройденных состояний

Для учета последовательности состояний использовался метод обучаемых предсказывающих фильтров. Этот метод основан на идее о применении обучаемых предсказывающих фильтров в качестве прототипов распознающей системы, предназначенной для классификации случайных процессов (4). Предполагается, что если предсказывающий фильтр был обучен на процессах одного класса, то его точность предсказания будет наивысшей для процессов только этого класса (процессы с другими статистическими характеристиками будут предсказываться с меньшей точностью). В данном случае, по существу, вводится новая мера сходства процессов, определяемая точностью предсказания текущей реализации процесса. Каждая входная реализация относится системой к классу процессов, который был использован для обучения фильтра, дающего для наблюдаемой реализации наилучшую точность предсказания.

Задачу классификации изменяющихся во времени или движущихся объектов с учетом порядка следования состояний можно свести к задаче классификации многомерных процессов. Как и ранее будем считать, что в процессе наблюдения за объектом представляющая точка обязательно находится в одной из областей пространства признаков. Другими словами, факт принадлежности представляющей точки к одной из об-

ластей пространства является одним из множества единственно возможных и несовместимых событий $X_1, X_2, \dots, X_N$

В общем случае события $X$ связаны в бесконечно усложняющуюся марковскую цепь. Такие цепи характеризуются переходными вероятностями $P(S_1, S_2, \dots, S_n, S_{n+1})$ такими, что вероятность события $X_j$ в $(n+1)$-ом наблюдении зависит от результатов всех предшествующих наблюдений $S_1, S_2, \dots, S_n$. В данном случае наблюдения обозначаются через $S_1, S_2, \dots$ и предполагается, что в каждом из проведенных наблюдений был получен определенный результат, т.е. наблюдалось какое-либо из событий $X_i$ ( $i = 1,2, \dots n$ ). Из определения событий следует, что

$$\sum_{i=1}^{m} P(S_1, S_2, \dots, S_n, S_{n+1}) = 1 \ . \qquad (32)$$

Переходные вероятности можно выразить как

$$P(S_1, S_2, \dots, S_n, S_{n+1}) = P[(S_1, S_2, \dots, S_n), (S_2, S_3, \dots, S_n, S_{n+1})] = P(a_n, b_{n+1}), \qquad (33)$$

где $a_n = (S_1, S_2, \dots, S_n)$; $b_{n+1} = (S_2, S_3, \dots, S_n, S_{n+1})$ $n = 1, 2, \dots$

причем $P(a_n, b_{n+1}) = P(S_1, S_2, \dots, S_n, S_{n+1})$,

если $a_n = (S_1, S_2, \dots, S_n)$ и $b_{n+1} = (S_2, S_3, \dots, S_n, S_{n+1})$;

если же $a_n = (S_1, S_2, \dots, S_n)$, а $b_{n+1} \neq (S_2, S_3, \dots, S_n, S_{n+1})$,

то $P(a_n, b_{n+1}) = 0$.

Совокупность вероятностей при различных значениях $n$ можно представить в виде матрицы переходных вероятностей $P_{a,b}$, составленной для бесконечного числа событий $a_n$ и $b_n$. Зная все элементы такой матрицы, можно при известной предыстории с некоторой вероятностью предсказать состояние объекта на любом этапе наблюдения $N_{n+k}$, выбрав для этого наибольшую вероятность для данного $k$. Переходные вероятности определяются в процессе обучения, т.е. в процессе длительного наблюдения за объектами.

Для каждого класса объектов составляется матрица переходных вероятностей и ее можно рассматривать как обученный предсказывающий фильтр, настроенный или обученный на определенный класс траекторий. В результате сравнения точности предсказания наблюдаемой траектории различными фильтрами (матрицами) можно вынести определенное решение о принадлежности данной траектории, а значит и наблюдаемого объекта к определенному классу. При этом наблюдаемый объект следует отнести к тому классу, переходная матрица которого дает наибольшие вероятности для тех событий, которые в действительности наблюдаются.

Практическое применение описанного метода затрудняется сложностью получения бесконечных переходных матриц. Решая практические задачи следует либо переходить к устойчивым бесконечно усложняющимся цепям, ассимптотами которых являются простые цепи Маркова, для которых

$$\lim_{n\to\infty} P(S_1, S_2, \ldots, S_n, S_{n+1}) \to P(S_n, S_{n+1}), \qquad (34)$$

где $P(S_n, S_{n+1})$ -вероятность перехода от результата, полученного в наблюдении $S_n$ к событию $X_i$ на $N_{n+1}$ наблюдении, либо необходимо использовать многосвязные цепи конечной сложности. Сложность матрицы тем больше, чем глубже взаимосвязь наблюдений.

Для исследования устойчивых многосвязных цепей бесконечной сложности, по отношению к которым справедливо соотношение (34), можно с некоторой погрешностью пользоваться всеми закономерностями простых цепей Маркова. Если же рассматривать цепи конечной сложности, то для выбранной глубины взаимосвязей следует построить конечную матрицу переходов, которая будет законом данной цепи,т.е. законом данного класса траекторий.

Пусть $X_1, X_2, \ldots, X_N$ единственно возможные и несовместимые события, появляющиеся в неограниченном ряде испытаний $S_1, S_2, \ldots$. Испытания соединяются в $\kappa$ -членные звенья ( $\kappa$ определяет глубину связи). Номер звена определяется номером первого его испытания, например,если

рассматривать звено, состоящее из $S_3$ , $S_4$ , $S_5$ , $S_6$ испытаний, то его номер будет 3. Обозначим звено испытаний с номером $h$ через $S_h^*$. Пусть $P(S_h, S_{h+1}, \ldots, S_{h+\kappa-1}, S_{h+\kappa})$ есть вероятность появления события $X_i$ в первом испытании после $\kappa$-членного звена с номером $h$. Введем $\kappa$-членные индексы $a$ и $b$ такие, что

$$P(a,b) = P(S_h, S_{h+1}, \cdots, S_{h+\kappa-1}, S_{h+\kappa}),$$

если $a = (S_h, S_{h+1}, \cdots, S_{h+\kappa-1})$; $b = (S_{h+1}, S_{h+2}, \cdots, S_{h+\kappa-1}, S_{h+\kappa})$

и $P(a,b) = 0$,

если $a = (S_h, S_{h+1}, \cdots, S_{h+\kappa-1})$; $b \neq (S_{h+1}, S_{h+2}, \cdots, S_{h+\kappa-1}, S_{h+\kappa})$.

Составив $N_\kappa$ значений этих индексов, получим матрицу вероятностей переходов, где событиями будут $a$ и $b$. Пользуясь формулами, приведенными в (3) можно сначала определить вероятность

$$P(S_h^* / X_i, X_{i+1}, \cdots, X_{i+\kappa-1})$$

появления последовательности событий в звене с номером $h$, а затем вероятность $P(S_h / X_i)$ появления события $X_i$ в испытании с индексом $h$. Практически можно использовать цепи с малой глубиной связи, так как при этом в одинаковой мере понизится точность предсказания для всех классов траекторий. Для примера рассмотрим случай, когда $\kappa = 2$.

Пусть $P(S_h^* / X_i X_{i+1})$ -вероятность появления событий $X_i$ и $X_{i+1}$ в звене с индексом $h$, а $P(S_h / X_i)$ -вероятность появления события $X_i$ в испытании с индексом $h$, т.е. первом испытании звена $S_h^*$. Результаты испытаний $S_1$, $S_2$, ..., $S_{n-1}$ известны. Требуется определить вероятность $P(S_h / X_i)$ появления события $X_i$ в испытании $S_h$. Пусть $a$ и $b$ двухчленные индексы, такие, что

$$P(a,b) = P(S_h, S_{h+1}, S_{h+2}); \qquad (35)$$

если $a = (S_h, S_{h+1})$; $b = (S_{h+1}, S_{h+2})$

и $P(a,b) = 0$ если $a = (S_h, S_{h+1})$; $b \neq (S_{h+1}, S_{h+2})$.

Тогда переходную матрицу можно представить в виде матрицы размерностью $N^2$, составленной из элементов $P(a,b)$.

Каждый из элементов этой матрицы определяет вероятность перехода, например, элемент матрицы $P_{111}$ равен вероятности появления события $X_i$ при условии, что в предыдущих двух испытаниях появлялось то же самое событие, т.е. $S_k = (X_i, S_{k+1})$, элемент $P_{132}$ равен вероятности появления события $X_2$ при условии, что в предыдущих двух испытаниях произошли события $X_1$ и $X_3$ и т.д.

Если определены все элементы переходной матрицы, то всегда по вероятности вида $P(a,b)$, известной на данный момент, можно определить вероятности вида $P^c(a,b)$, которые являются элементами новой матрицы, полученной путем возведения $\| P(a,b) \|$ в степень $C$, где $C$ - определяет номер того испытания, результат которого нас интересует. Эти вероятности можно подсчитать по формуле Перрона. И, наконец, интересующие нас вероятности определяются из отношения

$$P(S_c / X_i) = \sum_{\alpha} P(0, X_{i+f''}) P^{c-1}(a,b), \qquad (36)$$

где $a = (S_k = X_{i+f'};\ S_{k+1} = X_{i+f''})$; $b = (S_{k+1} = X_{i+f''};\ S_{k+2} = X_i)$;

$P(0, X_{i+f'}, X_{i+f''})$ - начальная вероятность, определяющая вероятность того, что в первой паре испытаний будут обнаружены состояния $X_{i+f'}$ и $X_{i+f''}$, где $f'$ и $f''$ - произвольные целые числа.

Таким образом, процедура классификации траекторий или, что то же самое, классификация объектов с учетом их предыстории методом обучаемых предсказывающих фильтров заключается в следующем: на основании наблюдений составляется переходная матрица для каждого класса траекторий ( обучение системы). Каждая такая матрица представляет собой обученный предсказывающий фильтр. В режиме классификации наблюдаемая траектория поступает на входе всех обученных фильтров и на каждом этапе наблюдения предсказывается будущее состояние траектории каждым фильтром в отдельности. В качестве предсказанного события выбирается такое, для которого соответствующая вероятность, полученная по формуле (36) для данной переходной матрицы максимальна. Наблюдае-

мая траектория относится к тому классу, фильтр которого дал лучшее предсказание. Наблюдение и предсказание продолжаются до тех пор пока точное предсказание будет на выходе только одного фильтра, после чего выносится решение о принадлежности наблюдаемой траектории.

Основная трудность в описанном методе состоит в определении переходных вероятностей. Чем короче звенья многосвязной цепи тем проще решается задача, но с другой стороны, чем длиннее звенья, тем точнее решается задача как предсказания, так и классификации.

## 4. Приложение

а) Пример применения критерия наиболее правдоподобных гипотез. На ЭВЦМ "МИНСК-22" был присчитан пример оптимального выбора наиболее правдоподобных гипотез в случае 3-х равновероятных классов. Обучающая выборка, состояла из 39-ти наблюдений по 13 на каждый класс, причем каждое наблюдение содержало 10 признаков. В период обучения были проведены вычисления вероятности распределения признаков в каждом классе $P(x_i/V_1)$; $P(x_i/V_2)$; $P(x_i/V_3)$,

а такие вероятности принадлежности очередного наблюдения к одному из трех равновероятных классов

$$P(X_y/V_1);\ P(X_y/V_2);\ P(X_y/V_3).$$

Далее определялись отношения $P_{max}/P^*_{max}$,

где $P_{max}$ наиболее правдоподобная гипотеза, а $P^*_{max}$ следующая за ней по абсолютной величине. В соответствии с критерием решение принималось при

$$\frac{P_{max}}{P^*_{max}} > A;$$

а так как константа A определяет вероятность ошибки

$$P_o = \frac{1}{A},$$

то по полученным результатам можно судить о зависимости вероятности ошибки на каждом этапе наблюдения. Эксперименты показали, что применение критерия наиболее правдоподобных гипотез позволяет уже при нескольких наблюдениях достиг-

нуть вероятности ошибки порядка 0,01, в то время как по другим известным критериям для множества гипотез требуется полный перебор всех гипотез при той же заданной вероятности ошибки.

б) Пример классификации движущихся объектов с учетом смены порядка следования состояний

Рассмотрим случай распознавания трех классов движущихся объектов с разбиением пространства признаков на четыре области, т.е. представляя описания объектов как конечный марковский процесс, принимающий только четыре состояния . Для составления матриц переходных вероятностей (вероятностей перехода траектории, описываемой движущимся объектом, из одной области пространства признаков в другую) в результате наблюдения этих трех классов составлена обучающая выборка. В примере классы движущихся объектов представлены моделями транспортных средств - легкового и грузового автомобилей, а также трактора.

Обучающие выборки для каждого класса состоят из результатов наблюдений одного и того же объекта в различных условиях ( изменение освещенности, различная установка объекта по ракурсу, изменение цвета, параметров телекамеры в процессе работы и т.д.) всего 40 x 11 реализаций для каждого класса. По обучающим выборкам вычислены математические ожидания траектории каждого класса, и, в соответствии с этими эталонами, пространство признаков по принципу равномерного распределения состояний разбито на четыре области. Каждая эталонная траектория представляется как марковский процесс конечной сложности, принимающий последовательно одно из четырех возможных состояний. Таким же процессом представляется каждая реализация обучающей выборки. Объекты каждого класса представлены одиннадцатью состояниями . Это объясняется тем, что при наблюдении движущегося транспорта выделяются только ракурсы заметно отличающиеся друг от друга (через $18^{0}$) при повороте на $180^{0}$. Вторая половина ракурсов представляет собой зеркальное отображение первой.

Далее по обучающей выборке для каждого класса вычисляется распределение вероятностей появления каждого состояния в отдельности, условные вероятности появления каждого состояния при условии, что в предыстории учитывалось одно, два, три состояния (одно-, двух- и трехчленные цепи Маркова) и составляются матрицы переходных вероятностей, которые и представляют собой обученные предсказывающие фильтры.

В режиме распознавания для каждой наблюдаемой траектории на каждом шаге (за шаг принимается ракурс объекта) предсказывается следующее состояние траектории. При совпадении предсказанного и наблюдаемого состояния хотя бы на одном фильтре распознавание заканчивается и траектория относится к тому классу в матрице которого вероятность перехода в это состояние максимальна. В том же случае когда по всем трем фильтрам предсказываемое и наблюдаемое значения не совпадают, предсказание продолжается, но уже с использованием предыстории на шаг больше и т.д.

В примере принято ограничение, вытекающее из реальной действительности, заключающееся в том, что наблюдение за движущимися объектами ведется последовательно, т.е. мы сначала видим нулевой ракурс, затем первый, второй и т.д. Принято, что траектория, описываемая объектом в процессе его движения, всегда начинается с нулевого отсчета, а не из середины или сзади (автотранспорт, как правило, не двигается по трассе боком или задним ходом).

При распознавании 120-ти реализаций, без учета предыстории (по первому наблюдению), средняя ошибка распознавания составила 65%, в то время как при распознавании с учетом предыстории на один шаг - 43%, два шага - 15%, а при учете трех шагов все траектории были классифицированы правильно.

## ЛИТЕРАТУРА


I. Вальд А. Последовательный анализ, Физматгиз, М.,I960.

2. Васильев В.И., Распознающие системы, Изд-во "Наукова думка", Киев, I969.

3. Васильев В.И., Реуцкий В.Е., Принцип учета предыстории входного сигнала при распознавании движущихся и изменяющихся объектов,"Автоматика", № 5, I968 г. (Часть I); "Автоматика",№ 3, I969 г. (Часть II).

4. Рогава С.Е., Распознавание случайных процессов с помощью сравнения точности действия ряда обучающихся предсказывающих фильтров,"Автоматика", № 5, I968 г.

## EELNEVA OLUKORRA ARVESTAMINE KUJUNDITE ERISTAMISEL

V.I. Vassiljev, V.E. Reutski
(Kiiev)

Resümee

Sageli tekib vajadus objekte eristada nende liikumise või arenemise käigus; iga olek on kirjeldatav tunnuste hulgaga; arengu käigus objekt kirjeldab teatavat trajektoori tunnuste ruumis. Ülesanne taandub trajektooride klassifikatsioonile.

Statistiliselt läheneb sellise ülesande lahendus järjendanalüüsi meetodi rakendamisele, kusjuures (igal sammul) tuleb kontrollida m + 1 erinevat hüpoteesi (objekt lugeda kuuluvaks mõnesse klassidest või jätkata vaatlusi).

Antud teineteist välistavate hüpoteeside rühma jaoks esitatakse järjendkriteeriumid ning tõestatakse teoreeme nende optimaalsuse kohta ( vaatluste arvu minimiseerimise mõttes).

Artikli teises osas taandatakse probleem mitmemõõtmeliste Markovi protsesside klassifitseerimisele.

Markovi protsessi kirjeldavad üleminekutõenäosused määratakse "õppimise" käigus. Iga objektide klassi jaoks moodustatakse üleminekumaatriks, mida võib vaadelda teatava trajektooride klassi jaoks häälestatud ennustusfiltrina. Iga objekt paigutatakse klassi,mille korral vastava trajektoori esinemise tõepärasus on suurim.

On kirjeldatud ka praktilist näidet.

## THE EVOLUTION OF OBJECTS IN PATTERN RECOGNITION

V.I. Vassilyev , V.E. Reutski
(Kiev)

Summary

There often arises necessity for object differentiation in the process of its motion and evolution; each state can be described by a set of variables; in its evolution the object describes a trajectory in the room of variables. The problem can be reduced to the classification of trajectories.

In statistics such solution is close to sequential analysis whereas at every stage the m + 1 hypothesis must be checked (the object can be classified or we must continue our observation).

For a given group of exclusive hypotheses sequential criteria are suggested and several theorems about the optimum of observations are proved.

The second part of the article deals with the classification of trajectories according to the classification of the Markov chains.

The probability of transition describing the Markov process is determined in the course of 'learning'. For each class of objects a transition matrix is drawn. Every object is put into a class where the probability of corresponding trajectory's occurence is maximum.

# ИНФОРМАЦИОННЫЕ СВОЙСТВА КОРОТКИХ ВЫБОРОК

В.И. Васильев

( Киев)

## I. Вступление

В работах (I), (2) уже указывалось, что те понятия классической теории информации, которые возникли на основе теории передачи сообщений по каналам связи (3), не могут полностью удовлетворить исследователей при решении некоторых специфических задач. Цель настоящей работы состоит в формализации таких интуитивных представлений, как обучающая информация, полезная и вредная обучающая информация, дезинформация, полезность и вредность дезинформации.

Пусть $S=\{S_i\}$ - конечное множество каких-либо событий. Природа событий $S_i$ в общем случае никак не ограничивается. Это могут либо различные действия какой-то системы, либо управляющие воздействия регулятора, либо возмущающие воздействия, либо различные поведения объектов, либо просто какие-то объекты. Множество $S$ состоит из подмножеств $V_1, V_2, \ldots, V_m$ таких, что $V_1 \cup V_2 \ldots \cup V_m = S$ и $V_i \cap V_j = \emptyset$ при $i \neq j$ и характеризуется распределением вероятностей $P(V)=\{P(V_1)=P_1, P(V_2)=P_2, P(V_m)=P_m\}$ где $P_i$ - вероятность события $S_i \in S$ такого, что $S_i \in V_i$. Будем считать, что в начальный момент распределения $P(V)$ нам неизвестно, а его оценка $Q(V)=\{q(V_1)=q_1, q(V_2)=q_2 \ldots q(V_m)=q_m\}$ осуществляется в процессе наблюдения за множеством $S$ в интервале времени $T << T^*$, где $T^*$ - интервал времени наблюдения, необходимый для достаточно точного определения вероятностей $P_i$, причем точность оценок, полученных по наблюдению в интервале времени $T$ никак не гарантируется. Серию наблюдений, произведенных в интервале времени $T$ будем называть короткой выбор-

кой. Если наблюдалась короткая выборка $Q_k$, то оценку распределения, полученную по этой выборке, будем обозначать через $Q_k(V)$.

2. Обучающая информация

Пусть кроме системы $S$ существует еще и система $S^*$. По каналу связи без помех ( $K$ ) информация о системе $S$ передается в систему $S^*$ (рис. I). В канале связи установлен прерыватель, который может управляться учителем ( $У$ ). Информация о системе $S$ накапливается в системе $S^*$ в виде оценок $Q_k(V)$. Систему $S^*$ можно интерпретировать, как наблюдателя, наблюдающего в определенные интервалы времени, за системой $S$, причем наблюдение может вестись непрерывно в течение интервала времени $T$, а может и прерываться по желанию учителя. Наблюдения формируют у наблюдателя некоторое представление о системе $S$, выражающееся оценкой $Q_k(V)$.

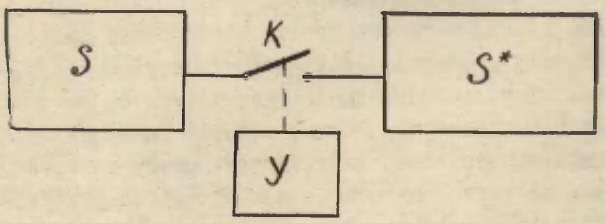
Будем называть энтропией системы $S^*$ (или энтропией наблюдателя) энтропию распределения $Q_k(V)$:

$$H(Q_k) = \sum_{i=1}^{m} q_i \log q_i. \qquad (I)$$

Очевидно, до начала наблюдения энтропия наблюдателя всегда будет максимальна и равна энтропии равномерного распределения. Действительно, до наблюдения у наблюдателя нет никаких оснований отдать предпочтение событиям из какого-либо подмножества $V_i$, так как он еще ничего не знает об этих событиях. Поэтому за начальный или нулевой уровень энтропии наблюдателя разумно принять энтропию равномерного распределения:

$$H_0 = \log m. \qquad (2)$$

Несомненно, что в процессе наблюдения представление наблюдателя о системе $S$ будет изменяться, т.е. распределение $Q(V)$ будет все ближе и ближе приближаться к распределению $P(V)$. (Пока речь идет только о канале связи без помех и без прерывателя, т.е. без учителя, ну и, конечно, предполагается, что распределение $P(V)$ отличается от равномерного). Значит будет изменяться и энтропия наблюдате-


S
K
S*
y

ля. Но изменить энтропию наблюдателя может только информация, поступающая по каналу связи. Поэтому разумно все то, что изменяет энтропию наблюдателя, т.е. изменяет его представления о системе $S$, назвать обучающей информацией и определить ее, как разность между начальной энтропией и энтропией наблюдателя после обработки выборки $Q_k$:

$$I(Q_k) = H_0 - H(Q_k). \qquad (3)$$

Изменение энтропии наблюдателя под воздействием наблюдаемой выборки можно рассматривать как процесс запасания информации в виде распределения вероятностей $Q_k(V)$. Энтропия наблюдателя, т.е. представление наблюдателя о системе $S$, зависит от той короткой выборки, которая наблюдалась. Если мы предположим, что содержание выборки может контролироваться (управляться) при помощи переключателя учителем, то и энтропия наблюдателя также может управляться учителем. Учитель почти всегда может подобрать такой поток обучающей информации, что система $S^*$ (т.е. наблюдатель) может принять любое желаемое состояние. Такой процесс можно назвать направленным обучением.

3. Противоречивость выборок и переобучающая информация

Предположим, что определена информативная мера $d(Q_1, Q_2)$ расхождения двух распределений $Q_1(V)$ и $Q_2(V)$. Пока мы не будем конкретизировать эту меру, так как еще нет оснований отдать предпочтение какой-либо известной мере подобного рода. Будем лишь считать, что эта мера обладает всеми свойствами расстояния, т.е.

$$\begin{aligned} &1.\ d(Q_1, Q_2) \geq 0; \\ &2.\ d(Q_1, Q_2) = 0 \rightleftarrows Q_1(V) = Q_2(V); \\ &3.\ d(Q_1, Q_2) = d(Q_2, Q_1); \\ &4.\ d(Q_1, Q_2) + d(Q_2, Q_3) \geq d(Q_1, Q_3), \end{aligned} \qquad (4)$$

и кроме того, если одно из распределений, например $Q_1(V)$, является равномерным, то

$$5.\ d(Q_1, Q_2) = I(Q_2) \qquad (5)$$

т.е. в этом случае мера расхождения $d(Q_1, Q_2)$ равна обучающей информации, необходимой для того, чтобы представление наблюдателя, определяемое распределением $Q_1(V)$, перешло в представление, определяемое $Q_2(V)$. Необходимость введения последнего требования будет пояснена несколько ниже.

Меру расхождения $d(Q_1, Q_2)$ можно рассматривать как информативную меру противоречивости выборок $Q_1$ и $Q_2$, если считать, что распределение $Q_1(V)$ получено в процессе наблюдения за выборкой $Q_1$, а распределение $Q_2(V)$ – за выборкой $Q_2$. Если же считать, что наблюдатель до наблюдения за выборкой $Q_2$, имел представление о системе $S$, определяемое распределением $Q_1(V)$, а после наблюдения приобрел представление, определяемое распределением $Q_2(V)$, то меру противоречивости (меру расхождения) $d(Q_1, Q_2)$ можно рассматривать как переобучающую информацию, содержащуюся в выборке $Q_2$ при условии, что наблюдатель до наблюдения имел представление о системе $S$, определяемое распределением $Q_1(V)$. Теперь понятно, что требование (5) определяет обучающую информацию, как частный случай переобучающей информации при условии, что до наблюдения наблюдатель ничего не знает о системе $S$, т.е. начальное распределение было равномерным.

## 4. Дезинформативность коротких выборок

Введенная мера противоречивости выборок помогает определить понятие дезинформативности выборки. Будем считать, что выборка, в результате обработки которой получено распределение $Q(V) = P(V)$ (т.е. оценки точно совпадают с действительными вероятностями событий), обладает нулевой дезинформативностью. Такую выборку, как и в математической статистике, будем называть представительной. Представитель-

ная выборка обладает нулевой дезинформативностью. Такое представление о дезинформативности совпадает с интуитивными представлениями. Действительно, если группа сообщений полностью отражает объективную реальность, то очевидно эта группа сообщений не несет в себе дезинформации.

В качестве меры дезинформативности конкретной выборки разумно принять меру противоречивости между представительной и рассматриваемой конкретной выборками:

$$D(Q_k) = d(P, Q_k). \qquad (5)$$

Смысловое содержание меры дезинформативности $D(Q_k)$ выборки $Q_k$ определяется количеством обучающей информации, необходимым для того, чтобы представление о статистических свойствах системы, определяемое распределением $Q_k(V)$ перешло в представление, определяемое действительным распределением. Мера дезинформативности дает информационную оценку тем заблуждениям или тем ложным представлениям, которые получает наблюдатель в результате статистической обработки конкретной выборки. Мера дезинформативности, как и мера противоречивости обладает всеми свойствами расстояния и характеризует собой уклонение полученного распределения от действительного.

## 5. Полезность обучающей информации и вредность дезинформации

Предположим, что накапливаемая в системе $S^*$ информация используется для решения какой-либо задачи $A$ при помощи класса алгоритмов $W$. Класс $W$-алгоритмов определяется правилом решения задачи, причем любое правило решения основано на использовании распределения $Q(V)$, полученного в результате обработки конкретной короткой выборки $Q_k$. Различные $W_k$- алгоритмы одного класса отличаются друг от друга только распределением $Q_k(V)$, т.е. зависят от короткой выборки $Q_k$, на основании которой получено распределение $Q_k(V)$.

Пусть $U = \{U_i\}$, $i = 1,2,\ldots,m$ – множество решений, которые могут быть приняты в процессе решения задачи $A$ классом $W$ алгоритмов. Пусть задана матрица штрафов $\Phi = [\Phi_{ij}]$, где $\Phi_{ij} = \Phi(U_i, U_j)$ – штраф за принятие решения $U_i$ при условии, что действительным решением является $U_j$, $j = 1, 2, \ldots, m$. Тогда средний риск можно определить по формуле:

$$R(A, W_k) = \sum_{i=1}^{m} \sum_{j=1}^{m} P(U_i U_j / A, W_k)\, \Phi_{ij}. \qquad (6)$$

где $R(A, W_k)$ – средний риск при решении задачи $A$ алгоритмом из класса $W$ с распределением $Q_k(V)$;

$P(U_i U_j / A, W_k)$ – вероятность того, что в условиях решения задачи $A$ алгоритмом $W_k$ будет принято решение $U_i$, тогда как действительным решением является $U_j$.

Средний риск является показателем качества решения задачи $A$ алгоритмом $W_k$.

Обозначим через $R(A, W_0)$ средний риск, полученный при решении задачи $A$ алгоритмом из класса $W$ с равномерным распределением $Q(V)$. Средний риск $R(A, W_0)$ определяет качество решения задачи $A$ в условиях полного отсутствия сведений о системе $S$.

Степень полезности информации, используемой для формирования $W_k$ – алгоритма, т.е. информации, содержащейся в короткой выборке, и, на основании которой получено распределение $Q_k(V)$, может быть определена как разность между показателем качества решения задачи $A$ в условиях полного незнания и показателем качества решения той же задачи после получения информации, содержащейся в короткой выборке

$$N(Q_k) = R(A, W_0) - R(A, W_k). \qquad (7)$$

Если эта разность положительная, то обучающая информация, содержащаяся в выборке $Q_k$ полезна для решения задачи $A$ алгоритмами из класса $W$. Если же эта разность отрица-

тельна, то нет никакого основания обучающую информацию выборки $Q_k$ называть полезной или хотя бы бесполезной. Для такой обучающей информации остается только одно название – вредная обучающая информация. Следует особо подчеркнуть, что понятие вредной обучающей информации является сугубо относительным и может употребляться только относительно конкретной короткой выборки, конкретной задачи $A$ и конкретного алгоритма $W_k$.

Для того, чтобы оценить количество полезной или вредной информации, предлагается использовать формулу:

$$I^*(Q_k) = N(Q_k) I(Q_k). \qquad (8)$$

Теперь обозначим через $R(A, W_p)$ средний риск, полученный при решении задачи A таким $W_p$ – алгоритмом, для которого оценки распределения точно соответствуют действительному распределению $P(V)$. Средний риск $R(A, W_p)$ определяет качество решения задачи $A$ в условиях полного знания статистических свойств системы $S$, т.е. точного знания распределения $P(V)$ (обучающая выборка представительна). Любая выборка, несущая в себе дезинформацию приводит к решениям, средний риск которых будет отличаться от $R(A, W_p)$. Поэтому степень вредности дезинформации, содержащейся в конкретной короткой выборке и в условиях решения конкретной задачи $A$ конкретным $W_k$ алгоритмом, можно оценить разностью:

$$W(Q_k) = R(A, W_p) - R(A, W_k). \qquad (9)$$

Эта разность может быть как положительной так и отрицательной. Если эта разность отрицательна, то дезинформация, содержащаяся в выборке $Q_k$ вредна для решения задачи $A$ алгоритмом $W_k$; если же эта разность положительна, то назовем эту дезинформацию полезной для решения задачи алгоритмом из класса $W$.

Само понятие полезной дезинформации кажется с первого взгляда неразумным. Но в своей практической деятель-

ности человек очень часто для достижения какой-либо цели пытается некоторые факты утаить, а некоторые усилить, т.е. представить почти обязательными. С точки зрения достижения поставленной цели, некоторое искажение представлений о системе событий может оказаться полезным. В народе говорят "не всякая правда - спасение, не всякая ложь - - погибель".

Для того, чтобы оценить количество дезинформации можно использовать формулу, подобную формуле (8)

$$D^{*}(Q_k) = \bar{N}(Q_k) D(Q_k). \qquad (10)$$

## 6. Приложения

### I. Информационные свойства коротких выборок в условиях решения задачи направленного обучения

Процессом направленного обучения будем называть такой процесс обучения, при котором информация об изучаемом объекте целесообразно управляется учителем. При этом предполагается, что канал связи, управляемый учителем, является единственным источником информации об изучаемом объекте.

Используя описанные ранее свойства коротких выборок можно поставить две практически интересные задачи:

I. Как, говоря правду и только правду, лучше всего обмануть "ученика"?

2. Как, говоря правду и только правду, подобрать эту правду для наилучшего достижения поставленной цели?

Степень обманутости ученика можно оценить сравнением сложившихся в результате направленного обучения статистических представлений об объекте с истинными его статистическими свойствами. При этом степень обманутости можно охарактеризовать мерой расхождения между действительным распределением и той оценкой, которая получена в результате направленного обучения.

Для того, чтобы решить первую из поставленных задач нужно организовать процесс обучения или, что то же самое, подобрать для обучения такую короткую выборку, которая несет в себе максимальную дезинформацию. Независимо от того, какая выбрана мера расхождения, она в любом случае является некоторым функционалом от действительного распределения и, от его оценки, т.е.

$$d(P, Q_k) = F(P, Q_k). \qquad (11)$$

Решим задачу нахождения экстремума-максимума функционала $F$ по $Q$ учитывая ограничения $\sum q_i = I$. Найденное значение $Q_{mx}(V) = \{q_1, q_2, \dots, q_n\}_{mx}$ является той целью, к которой должен стремиться учитель в процессе направленного обучения, решая первую задачу. Остается теперь организовать управление обучением так (или подобрать короткую выборку), чтобы в результате получить оценку $Q_{mx}(V)$. Это почти всегда осуществимо, кроме одного исключительного случая, когда требуется добиться оценки вероятности какого-либо события отличной от нуля тогда, когда его вероятность равна нулю. В таком случае можно достигнуть только субоптимального решения. Следует подчеркнуть, что формирование представлений ученика, определяемых оценкой $Q_{mx}(V)$, осуществляется только подбором показов реальных событий, характеризующих реальное поведение объекта, т.е. обучение ведется путем сообщения правды и только правды.

Вторая из поставленных задач может быть решена тогда, когда ученик, на основании полученных в результате обучения оценок, решает какую-либо конкретную задачу при помощи определенного класса $W$ -алгоритмов, причем качество решения этой задачи определяется средним риском, подсчитанным по формуле (6).

Средний риск является функционалом от $P(U_i U_j / A, W_k) = P(U_i U_j)$. В то же время

$$P(U_i U_j) = F^*(Q_k). \qquad (12)$$

Для решения поставленной задачи нужно, принимая во внимание ограничения $\sum q_i = 1$, найти такое $Q_{mm}$ , при котором функционал $R(A, W_{mx})$ достигал бы экстремума-минимума. При этом задача $A$ будет решаться алгоритмом из класса $W$ наилучшим образом. Для того, чтобы достигнуть этой цели, нужно организовать процесс обучения (подобрать короткую выборку) так, чтобы оценки , полученные в результате обучения, совпадали бы с $Q_{mm}$ . Здесь, так же как и в первом случае, при обучении используются сведения об изучаемсм объекте, содержащие правду и только правду, но эта правда подбирается так, чтобы получить оценки, совпадающие с $Q_{mm}$ .

## II. Информационные свойства коротких выборок в условиях решения задачи обучения распознаванию образов

В теории распознавания образов особое место занимает задача обучения. Если задачу распознавания рассматривать с точки зрения математической статистики, то обучение сводится к оценке распределения вероятностей $P(V_j/X)$ по наблюдению за обучающей последовательностью. ( $V_j$ - образы или классы, $X$ - описания изображений). После того как эти распределения каким-либо образом определены, $W$ - алгоритм, используя полученную оценку распределения, может дать ответ о принадлежности любого изображения из $S$ к тому или иному образу на основании сообщения о его описании $X$ .

Как и раньше будем считать, что $W$ - алгоритмы одного класса различаются между собой оценками распределения $P(V/X)$, т.е. распределениями $Q(V/X)$ , а $W$ - алгоритм различных классов отличаются правилом решения задачи распознавания.

В случае решения задачи распознавания каждому $W$ -алгоритму одного класса соответствует не одно распределение $Q(V/X)$ , а целый комплекс распределений, который может быть охарактеризован таблицей I. Каждый элемент в таблице $q_{ij} = q(V_j/X_i)$ соответствует вероятности появления изображения из образа $V_j$ при условии, что наблюдается описание $X_i$ , где $i = 1, 2, \ldots, \kappa$, $\kappa$ - число элементов множества $S$ .

Таблица 1

| | $x_1$ | $x_2$ | . . . . | $x_k$ |
|---|---|---|---|---|
| Q | $q_{11}$ | $q_{21}$ | . . . . | $q_{k1}$ |
| Q | $q_{12}$ | $q_{22}$ | . . . . | $q_{k2}$ |
| Q | $q_{13}$ | $q_{23}$ | . . . . | $q_{k3}$ |
| Q | . . | . . | . . . . | . . |
| Q | . . | . . | . . . . | . . |
| Q | . . | . . | . . . . | . . |
| Q | $q_{1m}$ | $q_{2m}$ | . . . . | $q_{km}$ |

Так как реальная система имеет конечное число состояний, то число возможных состояний можно сократить. Для этого как и в (I) можно ввести понятие неразличимости описаний. Описания $X_1$ и $X_2$ будем считать неразличными, если при появлении $X_1$ и $X_2$ распределение $Q$ остается неизменным, т.е. $W$ -алгоритм при распознавании этих сигналов будет пользоваться одним и тем же распределением и, естественно, будет отдавать предпочтение одной и той же гипотезе. Задавшись определенной точностью измерений можно все множество $X$ , состоящее из $k$-элементов отобразить во множестве неразличных сигналов $X^*$ , число элементов которого будет равно $t \leq k$. Естественно, что $q_{ij} > 0$ и $\sum_{i=1}^{t} q_{ij} = \sum_{i=1}^{k} q_{ij} = 1$. Таблица 1 с учетом неразличимости описаний не изменит своей структуры, кроме замены в последнем столбце индекса $k$ на индекс $t$ .

Если считать, что элементы таблицы определяются по коротким выборкам, то содержание таблиц, полученных по равным выборкам может быть различным, а значит и результаты решений одной и той же задачи распознавания $W$ -алгоритмами одного и того же класса тоже могут быть различными.

В общем случае, для $W$ -алгоритмов одного класса, вероятность общей ошибки распознавания можно получить по формуле

$$P_{on} = \sum_{\nu=1}^{L} P(X_\nu)[1 - P/W = V_j/X_\nu)], \qquad (13)$$

где $P/W = V_j/X_\nu$ - апостериорная вероятность образа, которой был указан $W$ -алгоритмом при появлении описания $X_\nu$ ,

$P(X_\nu)$ - априорная вероятность описания $X_\nu$ .

Назовем начальной неразрешимостью задачи распознавания $N(Q_0)$ для $W$ -алгоритмов вероятность ошибки, возникающей при использовании алгоритма $W_0 \in W$ , для которого все элементы таблицы I одинаковы.

Тогда степень полезности обучающей информации, содержащейся в выборке $Q_k$ можно определить по формуле

$$N(Q_k) = N(Q_0) - N(Q_k), \qquad (14)$$

где $N(Q_k)$ - неразрешимость задачи после обучения по короткой выборке $Q_k$ .

До обучения неопределенность ответов системы максимальна и соответствует равномерному распределению вероятностей ( все элементы таблицы I одинаковы). Эту неопределенность можно определить по формуле:

$$H_0 = \sum_{\nu=1}^{L} \sum_{j=1}^{m} P(X_\nu) P/V_j/X_\nu \log P/V_j/X_\nu = \log(m). \qquad (15)$$

После обучения по короткой выборке неопределенность ответов $H(Q_k)$ соответствует распределению $Q_k(V)$ и может быть определена по формуле, подобной (15).

Количество обучающей информации, полученной при обработке выборки $Q_k$, можно определить как

$$I(Q_k) = H_0 - H(Q_k) \qquad (16)$$

а количество полезной информации по формуле:

$$I^*(Q) = N(Q_k) I(Q_k). \qquad (17)$$

Каждое описание $\chi_j$ состоит из множества признаков, и каждый признак несет свою информационную нагрузку. Поэтому степень полезности и обучающую информацию выборки можно определять как для любой комбинации признаков, так и для каждого признака в отдельности, и если степень полезности какого-либо признака (комбинации признаков) будет отрицательна, то такой признак (комбинация признаков) является вредным при решении конкретной задачи распознавания при помощи конкретного $W$ -алгоритма и относительно конкретной обучающей выборки.

Если условия задачи требуют учитывать не только вероятность ошибки, но и функцию штрафов, то степень полезности должна подочитываться с учетом функции риска.

## ЛИТЕРАТУРА


I. Бонгард М.М. "Проблема узнавания", изд-во "Наука", Москва, 1967.

2. Стратонович Р.Л., О ценности информации, изд-во АН СССР, Техническая кибернетика, № 5, 1965.

3. Шеннон К. Работы по теории информации и кибернетики, изд-во иностранной литературы, Москва,1963.

4. Кульбак С. Теория информации и статистика, изд-во "Наука", Москва, 1967.

# VÄIKESTE VÄLJAVÕTETE INFORMATIIVSED OMADUSED

V. I. Vassiljev
(Kiiev)

**Resümee**

Artikli eesmärgiks on formaliseerida selliseid mõisteid nagu õppeinformatsioon, desinformatsioon, informatsiooni kasulikkus ning kahjulikkus.

Defineeritakse uuritav süsteem S ning "vaatleja" - süsteem S'. Õppeinformatsiooniks nimetatakse vahet

$$J(Q_k) = H_0 - H(Q_k)$$

kus $H(Q_k)$ on vaatleja entroopia peale väljavõtte $Q_k$ kasutamist, algentroopia $H_0 = \log m$, kus $m$ on klasside arv.

Kahe erineva väljavõtte $Q_1$ ja $Q_2$ jaoks määratletakse kauguse aksioome rahuldav erinevuse mõõt.

$$d(Q_1, Q_2)$$

ning konkreetse väljavõtte desinformatsioon defineeritakse kui vaadeldava väljavõtte $Q_k$ ning tegeliku jaotuse $P$ erinevuse mõõt

$$d(P, Q_k)$$

Ülesande $A$ algoritmi $W_k$ abil saadud lahendi headuse määramiseks kasutatakse keskmist riski $R(A, W_k)$ Väljavõttes $Q_k$ sisalduva informatsiooni kasulikkus $N(Q_k)$ määratletakse vaheks

$$N(Q_k) = R(A, W_0) - R(A, W_k),$$

kusjuures $R(A, W_0)$ vastab informatsiooni täielikule puudumisele süsteemi S kohta.

Desinformatsiooni kahjulikkust aga iseloomustab vastavalt vahe

$$\bar{W}(Q_k) = R(A, W_p) - R(A, W_k),$$

kus $R(A, W_p)$ vastab täieliku informatsiooni olemasolule süsteemi S kohta.

Artiklis näidatakse ka toodud mõistete rakendamist kujundite eristamise ülesandes.

# INFORMATIVE QUALITIES OF SMALL SAMPLES

V.I. Vassilyev
(Kiev)

Summary

The aim of the paper is to formalize such concepts as learning information, disinformation, the usefulness or harmfulness of information.

The object S and the "observer" S' are defined. The following difference describes the learning information

$$J(Q_k) = H_0 - H(Q_k)$$

where $H(Q_k)$ is the observer's entropy after the use of sample $Q_k$ ; the initial entropy is $H_0 = \log m$ , where $m$ is the number of classes.

For two different samples $Q_1$ and $Q_2$ a differential measure answering to distance axioms is suggested

$$d(Q_1, Q_2),$$

and the disinformation of a given sample is defined as the differential measure between the sample $Q_k$ under observation and the real distribution $P$ .

$$d(P, Q_k).$$

So determine the goodness of the solution of problem $A$ by the algorithm $W_k$ the mean risk $(A, W_k)$ is used.

The usefulness of information that is contained in sample $Q$ will be given in the difference

$$N(Q_k) = R(A, W_0) - R(A, W_k),$$

where $R(A, W_k)$ describes the total absence of information about system S.

The harmfulness of disinformation is characterized by the following difference

$$\bar{W}(Q_k) = R(A, W_p) - R(A, W_k)$$

where $R(A, W_p)$ describes the total information available about system S.

# РАСПОЗНАВАНИЕ ОБРАЗОВ НА ОСНОВАНИИ КАЧЕСТВЕННЫХ ПРИЗНАКОВ

Э. Тийт
(Тарту )

I. В социологических, экономических и медицинских исследованиях часто встречается проблема распознавания образов на основании качественных (номинальных) признаков, для которых не существует никакой упорядоченности (национальности, названия профессий, болезней и т.д.). Особенно сложной является эта задача в условиях полного отсутствия априорной информации: I) не известны образы ( обучение без учителя); 2) не известно исходное распределение.

В то же время необходимо решать такую задачу как стохастическую, т.е. найти для каждого элемента набор вероятностей о принадлежности этих элементов к определенным группам и принять окончательные решения о принадлежности элемента к какой-то группе на основании этого набора вероятностей.

В настоящей заметке дается алгоритм для решения таких задач, где для оценки вероятностей применяется распределение подходящим образом выбранного расстояния, которое имеет асимптотически нормальное распределение.

2. Рассмотрим пространство $R^k = R_1 \times R_2 \times \cdots \times R_k$, где $R_i$ – конечное множество точек:

$$R_i = \{a_1^i, a_2^i, \ldots, a_{s_i}^i\}, \qquad s_i \geq 2, \quad i = 1, 2, \ldots, k.$$

Для точек $a_i^i \in R_i$ не определено отношение упорядочения.

Точки пространства $R^k$ – $k$-мерные векторы $x = (x_1, \ldots, x_k) \in R^k$, где $x_i \in R_i$.

Пусть $i_p^r$ некоторое упорядоченное подмножество множества $\{1, 2, \ldots, k\}$:

$$i_p^r = \{i_1, i_2, \ldots, i_p\}, \quad p \leq k, \quad 1 \leq i_1 < i_2 < \ldots < i_p \leq k.$$

Подпространство $R^{i_p^r}$ пространства $R^k$ определяется как произведение

$$\prod_{i \in i_p^r} R_i = R_{i_1} \times R_{i_2} \times \cdots \times R_{i_p},$$



II

а вектор $\tilde{x}^{\psi_p} = ( x_{i_1}, x_{i_2}, \ldots, x_{i_p} )$ как проекция вектора $x$ на пространство $\tilde{R}^{\psi_p}$.

Проекции $x_i$ вектора $x$ в одномерные подпространства $R_i$ – компоненты вектора $x$.

Пусть $h = ( h_{i_1}, \ldots, h_{i_p} )$ – точка в пространстве $R^{\psi_p}$ а $A \subset R$ – произвольное множество. Тогда сечением множества $A$ в точке $h$ называется множество $A(h)$, определенное следующим образом:

$$A(h) = \{x;\ x \in A;\ x_i = h_i, \text{ если } i \in \psi_p\}.$$

Сечение пространства $R$ в точке $h$ определяется соотношением:

$$R(h) = \{x,\ x \in R;\ \text{если } i \notin \psi_p;\ x_i = h_i,\ \text{если } i \in \psi_p\}$$

а сечение $x(h)$ вектора $x$ в точке $h$ – соотношением

$$x(h) = (x_1, \cdots, x_{i_1-1}, h_{i_1}, x_{i_1+1}, \cdots, x_{i_p-1}, h_{i_p}, x_{i_p+1}, \cdots, x_n).$$

Расстояние в пространстве $R^n$ определяется как расстояние Хемминга:

$$d_i(x,y) = d(x_i, y_i) = \begin{cases} 0, & \text{если } x_i = y_i, \\ 1, & \text{если } x_i \neq y_i, \end{cases} \qquad (1)$$

$$d = \sum_{i=1}^{n} d_i(x,y). \qquad (2)$$

Ясно, что расстояние всегда удовлетворяет условию

$$0 \leq d \leq n.$$

3. Рассмотрим вероятностное пространство $(\Omega, \Sigma, P)$ и случайный вектор $X$, имеющий значения в пространстве $R$,

$$X(\Omega \Rightarrow R);\ X = (X_1, X_2, \cdots, X_n)$$

Вектор $X$ характеризуется своим распределением $P_X$; $P_X$ является вероятностной мерой в пространстве

$$\{R^n, \Sigma_R, P_X\},$$

где $\Sigma_R$ – множество всех подмножеств пространства $R_n$ (по нашему предположению это множество конечное).

Так как для $R_i$ ( $i = 1, 2, \ldots, n$ ) не определено упорядочение, то математическое ожидание и медиана распре-

деления не имеют смысла; характеристикой расположения для $X$ является его мода $c = (c_1, \ldots, c_k)$, которая определяется соотношением:

$$P_X(c) = \max_{y \in R^k} P_X(y). \qquad (3).$$

Классическим распределением, которым возможно сравнить описанное распределение, является равномерное. В рассматриваемом случае равномерное распределение определяется соотношением:

$$P(a) = \frac{1}{T}, \text{ для всех } a \in R^k,$$

где $T = \prod_{i=1}^{k} s_i$ - количество точек в пространстве $R^k$.

Концентрацию точки $k$ описывает коэффициент $\varkappa(k)$, определяемый равенством (4)

$$\varkappa_k(k) = T \cdot P_X(k). \qquad (4)$$

Ясно, что в случае равномерного распределения $\varkappa_k(a) = 1$, для всех точек $a \in R^k$. Для моды $c$ распределения $\varkappa_k(c) \geq 1$, и чем больше вероятность моды по сравнению с вероятностью равномерного распределения, тем больше значение $\varkappa_k(c)$.

Обычное определение мультимодальности не переносится на случай рассматриваемого пространства $R_k$, так как понятие локального максимума кривой распределения не имеет смысла. Поэтому вводим понятие псевдомоды следующим образом.

Назовем <u>псевдомодами</u> распределения $P_X$ все точки $c_i$ пространства $R^k$, при которых $\varkappa_k(c_i) > 1$ (см. (4)). Если распределение имеет моду (не является равномерным), то эта мода является псевдомодой.

Иногда может оказаться целесообразным применять следующее, более строгое определение:

Назовем точку $c_i$ <u>$K$-псевдомодой</u> распределения $P_X$, если имеет место неравенство

$$\varkappa_k(c_i) \geq K; \quad K \geq 1.$$

Понятно, что при каждом распределении существуют $K$-псевдомоды только для значений $K$, удовлетворяющих условию

$$1 \leq K \leq \varkappa_k(c) \leq T$$

($c$ - мода распределения).

Пусть $\nu_\rho \subset \{1,2,\dots,n\}$. Распределение проекции $X^{\nu_\rho}$ случайного вектора $X$

$$X^{\nu_\rho} = \{X_{i_1}, \dots, X_{i_\rho}\}$$

называется маргинальным распределением и обозначается $P_X^{\nu_\rho}$.

Распределением сечения $X(k)$ случайного вектора $X$

$$X(k) = (X_1, \dots, X_{i_1-1}, k_{i_1}, X_{i_1+1}, \dots, X_{i_\rho-1}, k_{i_\rho}, X_n)$$

является условное распределение $X$ при условии

$$X_{i_\nu} = k_{i_\nu} \quad (\nu = 1,2,\dots,\rho).$$

Понятие моды непосредственно переносится на случай маргинального и условного распределения.

Коэффициент концентрации определяется в случае маргинального распределения следующим аналогом равенства (4):

$$\varkappa_\rho(k) = T_\rho P_{X^{\nu_\rho}}(k), \tag{5}$$

где $T_\rho = \prod_{i \in \nu_\rho} s_i$.

В случае условного распределения соотношение (4) для коэффициента концентрации заменяется следующим аналогом:

$$\varkappa_\nu(k) = T_\nu P_{X(k)}(k), \tag{6}$$

где $\nu = n - \rho$, $T_\nu = \prod_{i \notin \nu_\rho} s_i$

4. Пусть задана выборка случайного вектора $X$

$$x_1, x_2, \dots, x_n,$$

где $x_j = (x_{j1}, \dots, x_{jn})$;

объем выборки $n$. На основании данной выборки определяется новая мера в пространстве $R^n$ – выборочное распределение $P_X^*$:

$$P_X^*(a) = \frac{h(a)}{n}. \tag{7}$$

где $h(a)$ – натуральное число, $0 \le h(a) \le n$, $\sum_{a \in R} h(a) = n$.

Для каждой точки $a \in R^n$ $P_X^*(a)$ является несмещенной и состоятельной оценкой для истинного распределения $P_X(a)$; при помощи $P_X^*(a)$ можно найти и оценки коэффициента концентрации $\varkappa_\nu^*(a)$ для каждой точки $a \in R^n$.

Для распределения $P_X^*$ так же, как и для распределения $P_X$ определяются маргинальные и условные распределения.

Выборочная мода $C^*$ определяется соотношением (3), где вместо $P_X(a)$ применяется $P_X^*(a)$ из (7).

Псевдомоды целесообразно определить, исходя из статистического метода сравнения данного распределения с равномерным распределением.

Именно, псевдомодой с уровнем значимости $\alpha$ (где $\alpha$ заданное достаточно маленькое число) называются такие точки $C_i^*$, при которых выполняется неравенство

$$P\left(P_X(C_i^*) \le \frac{1}{T}\right) < \alpha . \qquad (8)$$

Для вычисления вероятности в неравенстве (8) можно применять $\chi^2$-тест, а также эквивалентный с ним $t$-тест.

$\chi^2$-тест можно применять и для проверки гипотезы о равномерности рассматриваемого распределения. Статистикой для этого является сумма

$$\chi^2 = \frac{1}{Tn} \sum_{a_i \in R^k} \left(n - \bar{T}h(a_i)\right)^2, \qquad (9)$$

имеющая асимптотически $\chi^2_f$ распределение,

где $$f = \prod_{i=1}^{k} (s_i - 1) \qquad (10)$$

Формулы (8) - (10) сохраняются и в случае маргинальных и условных выборочных распределений.

5. Нахождение выборочных мод и псевдомод является теоретически весьма простой задачей: требуется только найти точку $a$, где $h(a) = max$.

Практически сравнение частот всех точек часто невозможно из-за большой размерности задачи.

В дальнейшем описывается алгоритм для приблизительного нахождения мод и псевдомод и сокращения размерности для задач, имеющих высокую размерность $k$ и большие количества точек на всех осях $s_i$.

Зафиксируем уровень значимости $\alpha$.

Рассмотрим все одномерные маргинальные выборочные распределения $P^{x}_{X_i}$ ( $i = I, 2, \ldots, \kappa$ ), определенные равенствами:

$$P^{x}_{X_1}(a_j) = \frac{h(a_1)}{n}, \quad j = 1, 2, \ldots, s_i.$$

Соответствующее равномерное распределение определяется равенством $P(a_j) = \frac{1}{s_i}, \quad j = 1, 2, \ldots, s_i.$

Найдем:

I) величину $\chi^2_{(1)}$ для всех координат $i$, определяющую отклонение от равномерного распределения (частный случай формулы (9)):

$$\chi^2_{(i)} = \frac{1}{s_i n} \sum_{j=1}^{s_i} [n - s_i h(a_j)]^2; \tag{I2}$$

2) вероятность

$$P_i = P(\chi^2 > \chi^2_{(i)}) \tag{I3}$$

из таблиц $\chi^2$-распределения ( $f_i = s_i - I$);

3) все значения $a_j$, для которых выполняется неравенство $x^i(a_j) \geq K$, где $x^i(a_j)$ вычисляется по маргинальному распределению $P_{X_i}$; это неравенство равноценно следующему:

$$h^i(a_j) \geq \frac{K_n}{s_i}; \tag{I4}$$

4) для точек $a_j$, удовлетворяющих условию (I4), найдем вероятности

$$P(P^i(a_j) \geq \frac{1}{s_j}) = 1 - P(P^i(a_j) < \frac{1}{s_j}) = q_{ij} \tag{I5}$$

исходя из $t$- или $\chi^2$-распределения.

Пусть $J$ - множество таких индексов $i$, при которых вероятность $P < a$ (см. (I3)). Рассмотрим далее подпространство $R^J$ пространства $R$; пусть его размерность $r$. По существу мы этим исключаем из дальнейшего осмотра те признаки, распределение которых довольно близкое к равномерному, как мало информативные.

Точки $a_j$, удовлетворяющие условию (I4), являются точками $a_{ij} \in R^{\kappa}$, находящимися на оси $R_i$ ($i = 1, 2, \ldots, \kappa$). Обозначим их множество буквой $A$ ($A \subset \tilde{R}^J$), и множество их маргинальных вероятностей $q_{ij}$ буквой $Q$.

Упорядочим величины $q_{ij}$ для каждого индекса $i \in J$:

$$q_{ij_1} \geq q_{ij_2} \geq \cdots \geq q_{ij_{n(i)}}, \tag{I6}$$

где $n(i) \geq 1$ из-за определения множества $J$; обозначим $q_{ij_1} = q_i$ и упорядочим величины $q_i$:

$$q_{i_1} \geq q_{i_2} \geq \cdots \geq q_{i_\kappa}. \tag{I7}$$

Последняя цепь неравенств определяет порядок осей в пространстве $R^J$, неравенства (I6) определяют порядок точек $a_{ij}$ на оси $R_i$, а условие

$$a_{i'j'} > a_{i''j''}, \text{ если } q_{i'} \geq q_{i''}$$

и упорядоченность всех точек множества $A$.

Выборочную моду $c = (c_1, \ldots, c_\kappa)$ в пространстве $\tilde{R}^J$ определим следующим образом:

$$c_{i_1} = a_{i_1 j_1}. \tag{I8}$$

Пусть $J_1$ – проекция множества выборочных точек ($x_1, \ldots, x_n$) в пространство $R^J$.

Эмпирическое распределение множества $J_1$ обозначим буквой $P_1$.

Образуем сечение $J_2$ множества $J_1$ в точке $a_{i_1 j_1}$; и рассмотрим маргинальное распределение множества $J_2$ на оси $R_{i_2}$; пусть это $P_2$. Найдем моду $a_{j_2}$ распределения $P_2$ и определим следующую координату выборочной моды

$$c_{i_2} = a_{i_2 j_2}. \tag{I9}$$

Пусть у нас определено $r$ координат выборочной моды, и

$$c_{i_r} = a_{i_r j_r} \tag{20}$$

Продолжим следующим образом: определим сечение множества $j_r$ в точке $a_{i_r g_r}$ ; пусть это будет множество $j_{r+1}$, а маргинальное распределение этого множества на оси $c_{i_{r+1}}$ обозначим через $P_{r+1}$ ; мода этого маргинального распределения $a_{g_{r+1}}$ и определяет $c_{i_{r+1}}$ -ую компоненту моды $c$ :

$$c_{i_{r+1}} = a_{i_{r+1} g_{r+1}}$$

Такой процесс определения продолжается до нахождения $c_{i_{k'}} = a_{i_{k'} g_{k'}}$ , так как тогда определены все координаты выборочной моды $c$ .

Вводим следующие обозначения:

$$\{a_{i_r g_r} : r = 2, \dots, k'\} + \{a_{i_1 j_1}\} = B_1$$
$$A \setminus B_1 = A_1$$

и сохраним в множестве $A_1$ исходное упорядочение. Выбираем первую точку из множества $A_1$ , пусть это будет

$$a_{k_1 h_1}, \qquad k_1 \neq i_1.$$

Определим новый порядок координатных осей $R_i \ (i \in J)$ следующим образом, исходя из упорядочения (I7). На первое место поставим $k_1$ , следуют индексы $i_1, i_2$ в порядке (I7) до места $k_1$ , дальше сохраняется (I7).

Применяя новое упорядочение координатных осей и новую исходную точку $a_{k_1 h_1}$ ,продолжаем все рассуждения определения моды $c$ , и получим в результате моду $c^1$ соответственно с координатами

$$c^1_{k_1} = a_{k_1 h_1}, \quad c^1_{i_1}, c^1_{i_2}, \dots, c^1_{i_{k'}}.$$

Одновременно определяются и множества

$$B_2 = \{a_{i_r h_r} : r = 1, 2, \dots, k'\},$$
$$A_2 = A_1 \setminus B_2,$$

и если $A_2 \neq \phi$, то можно определить моду $c^2$ , на основании первого элемента множества $A_2$ .

Таким образом получается конечное множество мод $c = c^0$, $c^1$ , ... , $c^v$ , притом

$$A_v = A_{v-1} \setminus B_v = \phi.$$

Для каждой моды в ходе ее вычисления легко находится и ее вероятность ( в смысле маргинального распределения пространства $\overset{*}{\times}$ )

$$\tilde{P}(c^i), \quad i = 0, 1, \dots, v.$$

При помощи этой вероятности можно проверить, какие из найденных мод являются $K$ -псевдомодами при заданном $K$, и какие являются псевдомодами соответственно заданному уровню $\alpha$.

6. Пусть $g=(g_1,\dots,g_n)$ некоторая фиксированная точка. Найдем расстояние выборочной точки $x_j$ до точки $g$:

$$d(g,x_j)=\sum_{i=1}^{n} d_i(g_i,x_j) \qquad (21)$$

Так как точка $x_j$ выборочная, то расстояние (21) является случайным. Найдем распределение этой случайной величины.

Величина $d_i(g,x_j)$ может иметь только значения 0 и 1; значит, имеет т.н. распределение Бернулли. Обозначаем

$$P(d_i(g_i,x_j)=1)=p_i,$$

тогда

$$1-p_i=q_i,$$

$$E\,d_i(g,x_j)=p_i,\quad D\,d_i(g,x_j)=p_iq_i,$$

и

$$E\,d(g,x_j)=\sum_{i=1}^{n}p_i,$$

$$D\,d(g,x_j)<n.$$

Значит, если величины $d_i(g,x_j)$ являются независимыми или достаточно слабо зависимыми, то на основании центральной предельной теоремы можно утверждать, что $d(g,x_j)$ имеет асимптотически нормальное распределение $N(\mu,\sigma)$, где

$$\mu=\sum_{i=1}^{n}p_i, \qquad (22)$$

$$\sigma^2=D\sum_{i=1}^{n}d_i(g,x_j)\approx\sum_{i=1}^{n}p_iq_i. \qquad (23)$$

Но для величин $p_i$ и $q_i$ легко найти выборочные оценки:

$$p_i=\frac{k(i)}{N},\quad q_i=1-p_i \qquad (24)$$

где $k(i)$ – число точек $x_j$ выборки, при которых имеет место равенство

$$d_i(g,x_j)=1.$$

7. Самую простую задачу классификации можно сформулировать при помощи $K$ -псевдомод следующим образом.

Пусть $g_1, g_2, \dots, g_m$ - множество $K$ -псевдомод. Найдем для каждой точки $x_i$ расстояния

$$d_{ij} = d(x_i, g_j).$$

Точку $x_i$ считаем принадлежащей к группе $G_{j^*}$, тогда и только тогда, если

$$d(x_i, g_{j^*}) = \min_j d(x_i, g_j).$$

В случае, когда найдется $l$ индексов $j_l^*$ таких, что

$$d(x_i, g_{j_1^*}) = d(x_i, g_{j_2^*}) = \dots = d(x_i, g_{j_l^*}),$$

то найдем числа [I] элементов в каждой из групп $G_{j_1^*}, \dots, G_{j_l^*}$. Пусть это числа $n_{j_1^*}, \dots, n_{j_l^*}$. Найдем $\max n_{j_k^*} = n_{j_s^*}$, и считаем $x_i$ принадлежащей к группе $G_{j_s^*}$.

Если максимальных чисел более одного, то выбирается группа с минимальным индексом.

8. Применяя введенные понятия можно сформулировать и более сложную, но зато более гибкую задачу классификации объектов (точек выборки) $x_j$ в пространстве $X_j$.

Пусть у случайного вектора $X$ имеется $m$ $K$ -псевдомод $g_1, g_2, \dots, g_m$. Считаем каждую моду центром группы $G_l$ $(l = 1, 2, \dots m)$.

Классифицируем элементы $x_i$ в группы $G_0, G_1, \dots, G_m$ следующим образом:

1. $x_j \in G_0$, если $P(x_j \in G_l) < E$ для $l = 1, 2, \dots, m$, (25)

2. $x_j \in G_l$, если $P(x_j \in G_l) \geq E$ и (26)

$$P(x_j \in G_l) \geq C \max_{r \neq l} P(x_j \in G_r). \quad (27)$$

3. $x_j \in \bigcup G_r$, если выполняется (25) для всех $r \in Q$, но не найдется такого индекса $l$, при котором выполняется (26).

---

[I] Рассматриваемый элемент $x_i$ не включается в рассматриваемые числа.

4. Если $P(x_j \in G_\ell) = P(x_j \in G_r)$,
то $x_j \in G_\ell$, если $P(g_\ell) > P(g_r)$, , (28)
и если $P(g_\ell) = P(g_r)$,
то $x_j \in G_\ell$, если $\ell < r$ (29)

Группы, определенные правилами I. – 3., являются в основном сферическими. Иногда является естественным объединить сферические группы в "сгущения" или плеяды – т.е. в несферические группы. Для этого дается правило 5.

5. Если выполняется одно из неравенств

$$P(g_\ell \in G_r) \geq H$$

или

$$P(g_r \in G_\ell) \geq H, \qquad (30)$$

то $G_\ell$ и $G_r$ объединяются в одну группу.

Рассматриваемая задача определяется 4-мя свободно выбранными параметрами $K$, $E$, $C$ и $H$. Рассмотрим их возможные значения.

Параметр $E$ регулирует число элементов $x_j$, не включенных в группы $G_1 - G_m$; если выбрать $E = 0^1$, то все элементы без исключения классифицируются в группы $G_1 - G_m$ и группа $G_0$ останется пустой. Понятно, что всегда $E$ должна быть достаточно малым числом (например, $E < 0{,}25$).

Параметр $C$ регулирует число элементов $x_j$, принадлежащих в более чем одну группу. Если взять $C = I$ и соединить с правилами $I^\circ - 3^\circ$ еще $4^\circ$, то получается задача, в которой каждый элемент принадлежит не более чем в одну группу.

Значит, при выборе параметров $E = 0$ и $C = 1$, набор правил $1^\circ - 4^\circ$ определяет т.н. однозначную и полную классификацию в сферические группы.

Число сферических групп определяется параметром $K$, который приобретает значения $K \geq I$.

При значении $K > \varkappa_r(C)$ (см. (3), где $C$ – мода распределения ) групп не образуется. При $K = \varkappa_r(C)$ образуется только одна группа (если у распределения существует $r$ мод, имеющих равную вероятность и, значит концентрацию, то групп будет $r$ ).

При значении $K = 1$ число групп теоретически может достигать значение

$$\min(n, T) - 1,$$

но в некоторых случаях даже при $K = 1$ число $K$-псевдомод может быть достаточно малым ( это случай сильно концентрированных распределений).

Параметр $H$ ($0 \le H \le 1$) характеризует интенсивность объединения исходных групп. Если выбрать $H = 1$, то объединения исходных групп не происходит. При $H = 0$ все исходные группы объединяются в одну группу. Подходящее значение параметра $H$ зависит от конкретной задачи.

9. При практическом решении выше формулированной задачи на первом шаге возникает вопрос: как определить вероятности, применяемые в условиях (21) - (26).

Предполагаем, что $P_X$ нам неизвестно. Применение выборочного $\kappa$-мерного распределения $P_X^*$ означало бы логический круг, если классифицируются те же элементы, на основании которых определялось распределение. Разбиение выборки на две части для оценки параметров и для классификации обычно невозможно или не целесообразно.

Одним возможным выходом из положения является применение распределения расстояния $d(g_\ell, x_j)$, которое, как выяснялось выше, апроксимируется нормальным распределением.

Предполагаем, что элементы группы $G_\ell$ имеют расстояние до моды $g_\ell$ :

$$d(g_\ell, x_j),$$

которое распределено $N(\mu^\ell, \sigma^\ell)$.

Тогда определим

$$p(y \in G_\ell) = 1 - F_\ell(d(g, y)) \qquad (31)$$

где $F_\ell$ - функция распределения $N(\mu^\ell, \sigma^\ell)$, и $y \in R^\kappa$.

Параметры $\mu^\ell$ и $\sigma^\ell$ определяются как выборочные по формулам (22) - (24), применяя при этом только элементы, принадлежащие à priori $x_\ell$ в группу $G_\ell$, значит,

$$p_i^\ell = \frac{1}{n_\ell} \sum_{r=1}^{n_\ell} d_i(g_\ell, x_r^\ell). \qquad (32)$$

I0. Алгоритм классификации реализуется в следующих этапах:

I$^0$. Определение $K$-псевдомод - центров $g_\ell$ групп $G_\ell$.

2$^0$. Априорное определение групп.

3$^0$. Вычисление параметров распределения на основании априорного распределения групп ( формулы I0$'$, I0$''$ и I8 ).

4$^0$. Определение вероятностей принадлежности к группам $G_\ell$ для всех элементов $x_i$ (формула (32).

5$^0$. Перераспределение групп на основании правил пункта 6.

6$^0$. Уточнение параметров распределения на основании новых групп $G_\ell$ ( повторение пункта 3$^0$).

Далее повторяются пункты 4$^0$ - 6$^0$, 3$^0$ и т.д., пока на очередном шагу перераспределения (5$^0$) ни один элемент не перемещается.

Докажем, что такой алгоритм конечный и процесс перераспределения сходится.

Пусть элемент $x$ переходит на некотором $\kappa$-том шагу из группы $G_\ell$ в группу $G_j$. В таком случае

$$d^\kappa(g_\ell, x) > \ldots > \tilde{\sigma}_\ell^\kappa$$

но тогда $\tilde{\sigma}_\ell^{\kappa+1} < \tilde{\sigma}_\ell^\kappa$

и тем более

$$d^{\kappa+1}(g_\ell, x) > \tilde{\sigma}^{\kappa+1}_\ell;$$

значит, элемент $x$ не может вернуться в группу $G_\ell$ на следующем и тем более на $(\kappa+m)$-ом шагу, $m > 1$. Так как число групп конечное, то и число нужных шагов конечное.

# KUJUNDITE ERISTAMINE KVALITATIIVSETE TUNNUSTE ALUSEL

E. Tiit
(Tartu)

Resümee

Rakendustes esineb sageli kujundite eristamise probleemi kvalitatiivsete (nominaalsete) tunnuste alusel.

Käesolevas artiklis esitatakse statistikute süsteem, mis lubab analüüsida objekte kvalitatiivsete tunnuste alusel: mood, pseudomoodid, kontsentratsioon; etaloonjaotuseks seda liiki ruumis soovitatakse ühtlast jaotust. Antakse ka algoritmid vastavate statistikute määramiseks küllalt kõrge dimensiooniga ruumis ning kriteeriumid mõningate statistiliste hüpoteeside kontrollimiseks.

Kaugus objektide $x$ ja $y$ vahel defineeritakse nn. Hemmingi kaugusena

$$d(x,y) = \sum_{i=1}^{n} d_i(x,y),$$

$$d_i(x,y) = \begin{cases} 0, & \text{kui } x_i = y_i \\ 1, & \text{kui } x_i \neq y_i \end{cases}$$

ning näidatakse, et küllaltki laiadel eeldustel on see kaugus asümptootiliselt normaaljaotusega.

Sellega on esitatud põhilise karakteristikute süsteem mitmesuguste statistikaülesannete, sealhulgas ka stohhastiliste kujundite eristamise ülesannete lahendamiseks kvalitatiivsete (mittejärjestatavate) tunnuste ruumis.

# PATTERN RECOGNITION BASED ON QUALITATIVE VARIABLES

E. Tiit
(Tartu)

Summary

In practise we often meet pattern recognition problems based on qualitative variables.

In the present paper a system of statistics is suggested that allows to analyse objects on the basis of qualitative variables: mode, pseudomodes, concentration, etc.; the standard distribution for such variables is homogeneous. Some algorithms for determining corresponding statistics in high dimensional room $R^{\omega}$ and several criteria for testing some of the statistical hypotheses are presented.

The distance between objects $x$ and $y$ is defined as Hemming's distance

$$d(x,y) = \sum_{i=1}^{\omega} d_i(x,y)$$

$$d_i(x,y) = \begin{cases} 0, & \text{if } x_i = y_i \\ 1, & \text{if } x_i \neq y_i \end{cases}$$

the distance is proved to have asymptotically normal distribution in many practical cases.

Thus we have dealt with the system of fundamental characteristics for solving statistical problems, amongst them stochastic pattern recognition problems in the room of qualitative (non-ranking) variables.

# ПРИНЦИПЫ КЛАССИФИКАЦИИ КАК ОДИН ИЗ ЭТАПОВ ФОРМАЛИЗАЦИИ ЯВЛЕНИЯ

Д.Х. Креймер

(Тула)

Системно-структурный подход как один из методов анализа явления позволяет четко выявлять место конкретного явления в более общей системе явлений. Те или иные конкретные данные могут быть отражены в схеме, построенной по принципу "от части к целому": например, объединим имеющиеся показатели в группы по три и каждую образованную таким образом группу обозначим собственным названием. Затем можем по этому же принципу объединять полученные группы по три вместе, давая им в свою очередь новое наименование. Таких иерархических уровней может быть много. Назовем их уровнями обобщения I-ой, 2-ой и так далее степеней. Нулевым уровнем обобщения называется исходный перечень факторов. Факторы четко определяются и имеют самостоятельные названия. На I-ом уровне обобщения располагаются категории I-ого уровня обобщения, в которые входят определенные группы исходных факторов. На каждом уровне обобщения располагаются соответствующие категории с самостоятельными названиями. Мы можем остановиться на любом уровне обобщения.

Классификация (распределение множества предметов на классы по общему для каждого класса признаку) и таксономия (наука, описывающая и распределяющая в группы организмы по видам, родам, семействам, отрядам, классам и типам) предполагает два различных подхода: I) Синтез: в группе классифицируемых объектов разыскивают общие элементы, показатели, один из которых и кладут в основу классификации. Это один из приемов статистической группировки. 2) Анализ: в данной системе стараются выделить такие основания классификации, которые позволили бы разложить систему на элементы, причем элементы при этом должны удовлетворять выдвинутым целью классификации критериям. Поэтому сам процесс анализа слага-

ется из постановки цели, для которой производится классификация, определения критериев, удовлетворяющих данной цели, и перебора различных оснований классификации, позволяющих наиболее приблизиться к выдвинутым критериям.

Рассматривая изучаемое сложное явление как систему взаимосвязанных элементов, требуется четко определить понятия системы, ее элементов, структуры, связи и функций этих элементов.

В.А. Лефевр [I] выдвигает в качестве основных критериев системы

I) наличие идеального и неизменного проекта,

2) возможность отклонения от этого проекта,

3) совершение системой действий по уменьшению этого отклонения,

4) системе не требуется питание энергией.

То есть системой называется подверженная изменению совокупность взаимосвязанных элементов, стабилизирующаяся за счет собственной энергии.

Элемент как составная часть сложного целого есть основная составляющая, из которых складывается система. Для своего определения элемент требует четкого критерия или основания классификации. Разные основания классификации элементов порождают разные системные представления.

Структура как определенная взаимосвязь, взаиморасположение элементов системы определяется основаниями классификации элементов и отношений. В иерархической структуре должно быть выдержано единство оснований классификации элементов (факторов и категорий) и отношений между иерархическими уровнями (уровнями обобщения).

Анализ явления возможен путем разложения его на составляющие, то есть на признаки, как анализируется, например, структура административного подчинения (субординации). Назовем это анатомией явления. Другой путь - разложение явления на вызывающие его факторы. Получается как бы генетическая, физиологическая структура. Назовем это физиологией явления. Физиологическое древо можно строить для каждой вершины ана-

томического древа. Физиологическое древо отражает причинно-следственные зависимости между факторами, векторы влияния их друг на друга, что открывает возможность математического моделирования. Анатомическое древо характеризуется отношениями между признаками типа общее-частное, целое-часть, род-вид . Ю. И. Клыков [2] выделил 206 видов отношений, 50 из которых являются основными и позволяют описать все многообразие существующих взаимоотношений, правда, в основном лишь технических объектов (вещь-держатель, вещь-место и т.д.). Для социальных объектов этот список надо дополнять, что относится к компетенции специалистов по математической лингвистике.

Если предприняты попытки классификации наук, то логично продолжить это направление до классификации категорий каждой науки, добиваясь предельно однозначной и математической трактовки каждой категории. Задавшись стандартными категориями и стандартными отношениями между ними (число которых ограничено), можно описать любую ситуацию в стандартных терминах конкретной науки. Причем ситуация будет описана на желаемом для нас уровне обобщения. Задача фактически сводится к выработке системы категорий , отражающих все оттенки изучаемого явления. Для науковедения нужен свой Линней, систематизирующий категории и подкатегории каждой науки в виде стройной иерархии, и свой Дарвин, раскрывающий законы эволюции этой структуры,древа познания бытия.

Категория, как и закон, есть результат обобщения, выявления сущностного, устойчивого. Но если закон отражает определенную существенную связь, то категория есть лишь названия объектов, между которыми устанавливается связь, Закон отражает характер связи категорий. Категория же отражает понятия разной степени общности в промежутке общее-частное, род-вид.

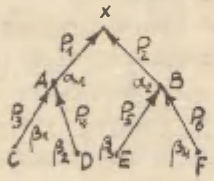
Связь элементов в системы характеризуется вектором влияния одного элемента на другой ( то есть точкой приложения, направлением и величиной). Рассматривая элементы

системы как относительно независимые, характеризующиеся своим набором атрибутивных признаков ( то есть в конечном счете как системы в системе), очень перспективен аксиологический подход ( аксиология - наука о ценностных отношениях) для анализа взаимосвязи [3] : значимость одного элемента системы для бытия другого (то есть для правильного функционирования системы этого элемента). Значимость (ценность) в значительной мере выражается причинностью,но не тождественная ей. Значимость удобно замерять по выводимым вышеприведенным способам "анатомофизиологическим" схемам-графам.

Функции элементов системы есть не что иное,как значимость данного элемента для конкретного объекта, которым может быть другой элемент, группа их, система в целом или любое основание классификации, сравнения. То есть функция предполагает объект, по отношению к которому определяется значимость элемента. Но можно выделить и так называемый содержательный носитель этой значимости, то есть те аспекты деятельности элемента, которые обусловливают или при помощи которых он оказывает влияние на объект. Тогда функции будут отражать способы, пути влияния элемента на объект. Можно различать частоту пользования различными путями влияния, значимость единовременного влияния через каждый из путей в отдельности и отсюда выводить среднестатистические (типичные) функции и значимость элемента для объекта.

Рассмотрим способы математической записи структур.Если разложить явление на вызывающие его факторы I-ой, 2-ой и т.д. ступеней, то есть построить ориентированный граф в виде дерева, то поскольку ребра графа указывают направление связей и можно измерить тесноту связи (значимость одного фактора для бытия другого), то можно рассчитать изменение конечной вершины дерева по изменению нижестоящих вершин. Если разложить явление на два вызывающих его фактора, то данную структуру можно изобразить графически :


и математически: $X = f(A,B)$ Поскольку в общественных процессах связь факторов корреляционная, а не функциональная, примем, что $A$ вызывает $X$ с вероятностью $P_1$, а $B$ – с вероятностью $P_2$. При наличии обоих факторов $X$ вызывается с вероятностью $P_1 + P_2$. Если $X$ – дискретная величина, качественная, то наличие и мера факторов $A$ и $B$ изменяет математическое ожидание появления или изменения $X$, а если непрерывная, то с изменением $A$ в $\alpha_1$ раз $X$ изменяется в $P_1 \alpha_1$ раз, а от изменения $B$ в $\alpha_2$ раз – соответственно в $P_2 \alpha_2$ раз. Причем здесь сделано допущение, что теснота связи численно равна вероятности. Можно двинуться далее по нашему дереву-графу.




Фактор $A$ вызывается факторами $C$ и $D$. Если изменить $C$ в $\beta_1$ раз, то $A$ изменится в $P_3 \beta_1$ раз и так далее. Для наглядности изобразим полученное дерево. Тогда на отрезке $C \to A \to X$ вероятность (теснота связи) между $C$ и $X$ $P_{CX} = P_3 P_1$. Подобным образом можно описать зависимость в многозвенной цепи, представив в конечном счете как $X = f(A,B,C,D,E,F) = f[\varphi_1(C,D), \varphi_2(E,F)]$ По связи между $C$ и $A$ и между $A$ и $X$ можно узнать связь между $C$ и $X$. Если известен метод, определяющий тесноту связи между факторами, и метод, определяющий направление этой связи (граф), то использование обоих методов позволит вычислить степень причинности. Таким образом появились рычаги управления оптимизируемой величиной $X$. Вне моделирования, то есть упрощенного представления явления для выяснения его механизма и последующего использования этого механизма для управления явлением, нет пути к оптимальному

планированию и управлению. Составление качественной схемы факторов в виде дерева-графа, позволяющей упорядочить большое их число, есть первый необходимый шаг в моделировании явления.

Математическое описание явления при помощи связанных с ним признаков позволит оптимизировать данное явление. Но можно поставить задачу более широкого класса, представив оптимизацию не как акт, а как процесс, продолжающийся более или менее успешно в зависимости от реакции противоположной стороны. Таким образом, введение временной координаты позволяет описать этот процесс в виде игры, например, с природой. После качественного анализа, то есть разложения явления на составные части, синтезируют их в модель взаимодействия этих частей. Для синтеза структур может быть использован аппарат теории вероятностей и математической статистики, корреляционный, дисперсионный, факторный анализ, последовательный анализ Вальда, цепи Маркова, теория графов, теория игр, динамическое программирование, разностные дифференциальные уравнения и т.д.

Одна из задач социологии - формализация социальной жизни, обоснование принятых формализованных структур, синтез выделенных формальных элементов в математической модели. Но для того, чтобы делать выводы из материала, подвергшегося той или иной формальной обработке, надо вникнуть в существо примененных формальных методов, иначе мы можем неадекватно интерпретировать полученные результаты. Поэтому повышаются требования к математической подготовке социологов.

Системно-структурный подход позволяет таким образом разложить по полочкам любое наисложнейшее явление, открывая пути для дальнейшей формализации выделенных и расклассифицированных элементов. Найдя место изучаемого явления в более общей системе и низводя всякий вопрос до уровня соотношения структурных элементов, можно решить его не только в качественной, но и в количественной форме. Выделив элементы (категории) и набор правил ( вида взаимосвязи между категориями), можно описывать не только со-

стояния, но и процессы, подлежащие нашему рассмотрению. Причем четкое определение категорий и правил позволяет представить их в формально-логическом и логико-математическом виде, позволяющем строить математические модели изучаемого явления или процесса.

Можно встретить возражения, что применительно к такой сложной системе, как общество людей, невозможно строить такие формальные системы, поскольку они не смогут правильно отразить все ее сложные взаимосвязи. Но этот агностицизм порожден страхом перед громадностью нерасчлененной классификацией проблемы. Расчленив ее и рассматривая более узкий вопрос, станет понятна относительная самостоятельность выделенных подсистем и в разрезе этой относительной самостоятельности такой подход правомерен.

Рассмотрим способы фиксации классификации с точки зрения наглядности, понятности содержания. Системность и наглядность имеют одинаковые цели: охватить явление в целом, только в одном случае аналитически, а в другом - визуально, образно. Синтез рационально-чувственных анализаторов оказывается очень плодотворным рычагом дальнейшего анализа явления. Графические методы подачи материала (схемы, графики, структуры) концентрируют в себе самую суть отображаемого явления, тенденцию, его составные части. Как правило, структура должна быть только графической. В то же время многие монографии о структуре сознания и других проблем исторического материализма ограничиваются лишь описательным материалом, в то время как описание в данном случае оправданно лишь для вывода и пояснения графической структуры, что позволит наглядно оценить ее содержательность и облегчит дальнейшую формализацию выведенной структуры.

Взаимосвязи явления удобно представлять в форме ленточных моделей, представляющих собой ориентированный граф в виде дерева. По сравнению с циклическими графами это позволяет четко указывать основание классифика-

ции в каждом разветвляющемся узле графа. При этом возможны повторения одних и тех же элементов системы, но зато видны основания классификации элементов системы и влияние каждого элемента на определенную сторону явления, поскольку сложное явление имеет много граней, могущих стать целью, объектом исследования, и много подходов, точек зрения на каждую грань. Введя таким образом четкие критерии, открываются пути для применения к ним различных методов численного анализа, применимость которого к сложным явлениям (в частности, к социальным) ограничена именно неразработанностью подобной предварительной формализации.

## ЛИТЕРАТУРА


I. В.А. Лефевр. Конфликтующие структуры. М.,"Высшая школа", I966.

2. Ю.И. Клыков. Модельный метод управления динамическими ситуационными средами. Диссертация , М., Московский энергетический институт,I967.

3. В.А. Василенко, Ценность и ценностные отношения. В кн.: "Проблема ценности в философии", М.- Л., "Наука", I966.

## KLASSIFITSEERIMISPRINTSIIP KUI ÜKS NÄHTUSE FORMALISEERIMISE ETAPPE

D.H. Kreimer
(Tuula)

Resümee

Artiklis käsitletakse süsteem-struktuurset lähenemist nähtuste analüüsimisel, kusjuures nähtust iseloomustavad näitajad ühendatakse rühmadesse. Hierarhiline rühmitamine annab mitme erineva tasemega üldistused.

Autor defineerib süsteemi, elemendi, struktuuri mõisted, selgitab seose ja faktori tähendust. Seejärel vaadeldakse struktuuride esitamise võimalusi graafide abil, mille (orienteeritud) kaared kujutavad faktorite vahelise seose mõju suunda. Niisuguse esituse eesmärgiks on peaasjalikult keerukate süsteemide näitlikustamine.

## THE PRINCIPLE OF CLASSIFIC..TION AS A STAGE IN PHENOMENA FORMALIZATION

D.H. Kreimer
(Tula)

Summary

A systematic-structural approach to phenomena classification is dealt with in this article. The features characterizing a phenomenon are united into groups. Generalizations on several levels can be made on the basis of hierarchical grouping.

The author defines the following concepts: a system, an element, a structure, explains the meanings of such concepts as a relation and a factor. Then some possibilities of expressing structures by mean of graphs are presented. The orientated arcs of graphs defict the trend of influence of interrelated factors. The aim of such treatment is mainly the graphic presentation of complex systems.

# ОБ ОДНОМ СПОСОБЕ КЛАССИФИКАЦИИ НА ГРАФАХ

И.Э. Муллат

(Таллин)

В настоящем сообщении рассматривается задача выделения "достаточно полных" графов-частей в заданном графе. Подобная задача в некотором смысле, по-видимому, похожа на задачу идентификации объектов. Усматривается нетривиальное содержание отношений между объектами выраженное в форме графа. Мы не претендуем на бесспорность высказанных утверждений и намеренно не будем торопиться с точной формулировкой понятия "достаточно полный" граф, указывая тем самым лишь на то, что объекты-вершины графа "близки" в какой-то мере.

В качестве примера можно привести граф, каждая из вершин которого обладает по крайней мере степенью $n/2$, где $n$ число всех вообще вершин графа, другой пример представляет, скажем, граф с числом ребер более 90% от наибольшего возможного числа ребер, т.е. числа ребер полного графа и т.д. В обоих приведенных примерах желательно, чтобы ребра "равномерно" заполняли граф.

Ниже мы везде рассматриваем неориентированные графы. Предполагается, что читатель знаком с элементарной терминологией теории графов хотя бы в объеме первой главы монографии Оре [1].

Указанные выше примеры, по-видимому, должны быть ясны с точки зрения того, какую цель мы ставим при идентификации частей в заданном графе. Из сказанного неявным образом явствует, что мы все же вынуждены измерять "полноту" графа в виде некоторых функций от элементов образующих граф. Условимся впредь называть эту меру весом. Для вершин, например, это может быть степень, для ребра, скажем, количество путей определенной длины ведущих из одного конца ребра в другой и т.д. Все дальнейшие рассуждения проводятся на множестве частей заданного графа $G$, хотя ничто не мешает нам рассматривать подобные конструкции на произвольном множестве.

Пусть $\{H\}$ множество частей графа $G$, $V(G)$ - множество вершин графа $G$, $\alpha = [a,b]$, $a$, $b \in V(G)$ - общий вид ребра. Основное содержание предлагаемой конструкции состоит в том, что предполагается наличие некоторой системы весов $\{\pi_H\}$ для каждой части графа $G$, причем существует следующая зависимость между системами весов различных частей:

для любого ребра $\alpha \in H$ и любого ребра $\beta \in H - \alpha$ выполняется

$$\pi_{H-\alpha}(\beta) \leq \pi_H(\beta). \qquad (I)$$

Иными словами в результате удаления ребра из части графа образуется новая система весов на оставшейся части, причем, удаление ребра влияет на веса оставшихся ребер только в сторону уменьшения. Сказанное поясним примером, где вместо ребер рассмотрим вершины, а весом будем считать степень вершины. Действительно, удаляя из графа любую вершину вес любой оставшейся может лишь уменьшиться. Ради удобства расшифровываем обозначение $\{\pi_H\}$ - система весов относительно части $H$.

Рассмотрим следующую функцию на множестве частей $\{H\}$ графа $G$:

$$f(H) = \min_{\alpha \in H} \pi_H(\alpha).$$

Мы предлагаем алгоритм нахождения такой части $H$ графа $G$, на которой $f(H)$ достигает максимума. Пусть

$$G = G_0 \cup G_1 \cup G_2 \cup \dots,$$

сумма по ребрам своих частей $G_i$.

Определяем последовательность

$$\Gamma_0, \Gamma_1, \Gamma_2, \dots \Gamma_k,$$

где $\Gamma_0 = G$, $\Gamma_{i+1} = \Gamma_i - G_i$. Последовательность считаем удовлетворяющей следующим свойствам:

а) вес $\pi_{\Gamma_i}$ любого ребра принадлежащего $\Gamma_i$, но не принадлежащего $\Gamma_{i+1}$ строго меньше $f(\Gamma_{i+1})$;

б) в $\Gamma_k$ не существует такой части $H$, чтобы выполнялось строгое включение $H \subset \Gamma_k$ и $f(\Gamma_k) < f(H)$, заметим, что $\Gamma_k = G_k$.

Лемма. Для любой части $H' \in \{H\}$ выполняется

$$f(H') \leq f(G_k)$$

и часть $G_k$ единственна с точностью до определяющей последовательности $\Gamma_i$.

Доказательство. Предположим, что существует часть $L$ графа $G$ отличная от $G_k$ такая, что

$$f(G_k) \leq f(L). \qquad (2)$$

По предположению не может быть, чтобы $L \subset G_k$, случай равенства исключен из рассмотрения по предположению, следовательно существует ребро $l \in L$ такое, что $l \notin G_k$. Пусть $l$ принадлежит некоторому $\Gamma_j$, очевидно, что $j < k$. Не умаляя общности можно считать, что

$$l \notin \Gamma_{j+1} \quad \text{и} \quad L \subset \Gamma_j.$$

Пользуясь свойством а), выводим неравенства

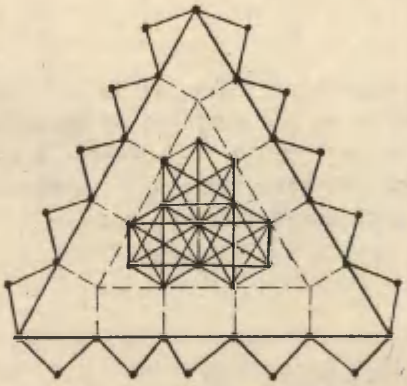
$$\pi_{\Gamma_j}(l) < f(\Gamma_{j+1}) \leq f(G_k). \qquad (3)$$

Пусть $m \in L$ и вес данного ребра минимален относительно всей части $L$, тогда вес ребра $l$ относительно части $\Gamma_j$ строго меньше веса ребра $m$ относительно части $L$. Но $L \subset \Gamma_j$, откуда используя неравенства (2) и (3) и вспоминая основное свойство (I) системы весов (удаление ребер), заключаем, что вес ребра $l$ относительно части $L$ должен быть строго меньше веса ребра $m$ относительно той же части, и получаем противоречие. Лемма доказана.

Примечание. Допустив существование различных последовательностей $\Gamma_i$, определяющих $G'_k$, легко установить равенство всех $G'_k$.

В заключение отметим, что алгоритм разбиения заданного графа $G$ на прямую сумму по ребрам удовлетворяющей свойствам а) и б) осуществлен практически на ЭВМ "Раздан-3". В качестве весов рассматривалась некоторая вероятностная мера. Время счета немногим превышало время решения системы уравнений (линейных) с $n$ неизвестными при объеме памяти $n^2$.

Укажем некоторые характерные черты работы вышеуказанного алгоритма на следующем графе:



Алгоритм позволил выделить части графа обведенным сплошной линией. Веса в процессе счета могли просчитываться с заданной точностью, что увеличивало мобильность программы в зависимости от характера задачи. В программу вводился такие особый параметр, позволивший укрупнять или сужать идентифицируемые части в графе.

ЛИТЕРАТУРА

I. О. Оре, Теория графов, "Наука", М., 1968.

ÜHEST KLASSIFITSEERIMISMEETODIST GRAAFIDEL

J. Mullat
(Tallinn)

Artiklis tuuakse ära klassifitseerimisalgoritm graafidel. Algoritmi kasutatakse elektronarvutil "Razdan 3".

THE CLASSIFICATION ALGORITHM ON THE GRAPHS

J. Mullat
(Tallinn)

Summary

In the paper a classification algorithm on the graphs is given.

The algorythm is calculated by the computer "Razdan 3".

# О ВЫДЕЛЕНИИ ИНФОРМАТИВНЫХ ХАРАКТЕРИСТИК В СОЦИОЛОГИИ

А.И. Тишин
(Фрунзе)

В социологии очень часто требуется определить на основе анкетного опроса общие и специфические черты, характеризующие, например, людей различных национальностей или различного возраста, или групп респондентов, которые по-разному выражают свои чувства, отношения к тому или иному событию и т.д. При этом нередко требуется из множества характеристик $S(s_1, s_2, \ldots, s_n)$ какого-либо явления или процесса выделить наиболее информативное подмножество их $X(x_1, x_2, \ldots, x_k) \subset S(s_1, s_2, \ldots, s_n)$, подобрав для этого определяющую множество $X$ функцию $\delta$ такую, чтобы "стоимость затрат" $R$ на решение поставленной задачи была минимальной. Эта задача распознавания образов по классификации Н.Г. Загоруйко [1], [2] относится к третьему типу.

Пусть имеется массив из $N$ анкет, причем каждая анкета состоит из $k$ вопросов (признаков), а каждый признак имеет то или иное количество возможных вариантов ответов (градаций). Ввиду того, что каждый признак со взаимопредполагающимися градациями можно представить в виде двух или более признаков со взаимоисключающимися градациями, то, не нарушая общности, полагаем, что каждый признак содержит только взаимоисключающиеся градации. Принадлежность $i$-ой градации $l$-ому признаку обозначим через $l(i)$, а количество градаций в $l(i)$-ом признаке - через $p_{l(i)}$. Очевидно, если $i_1, i_2, \ldots, i_m \in l$, то $p_{l(i_1)} = p_{l(i_2)} = \ldots = p_{l(i_m)} = p_l$. Общее число градации в анкете определяется как сумма градации всех признаков, т.е. $n = \sum_{l=1}^{k} p_l$. Для удобства пронумеруем все

градации от 1 до $n$ , а все признаки - от 1 до $k$ в том порядке, в каком они расположены в анкете.

Тогда исходный социологический массив можно представить в виде симметричной $n \times n$ матрицы $\{a_{ij}\}$ , где $a_{ij}$ - число людей, одновременно ответивших на $i$ -ю и $j$ -ю градации. Отождествим множество исходных характеристик $S$ со множеством всех градаций в анкете. Тогда поставленную задачу, предварительно оформив, можно решить так.

Выделяем в анкете совокупность интересующих нас градаций, обозначенную фиксированным признаком $l$ , например, признаком "национальность". Тогда градации этой совокупности на множестве всех респондентов "определяют" таксоны $i_1$ , например, "русские", $i_2$ -"киргизы", $i_3$ - "эстонцы" и т.д., наконец, $i_{p_l}$ -"другие национальности". Число таксонов при этом будет $p_l$ . Требуется из множества $n - p_l$ оставшихся характеристик-градаций выделить а) общие информативные характеристики для всех $p_l$ таксонов и б) специфичные информативные характеристики для каждого из рассматриваемых таксонов. Общими характеристиками, например, для представителей всех национальностей СССР могут быть любовь к своему народу, трудовой подъем в коммунистическом строительстве и т.д., а специфические характеристики могут определяться особенностями традиций, культуры, экономики и т.п. той или иной нации или народности.

Простейшее решение такой задачи дает следующий алгоритм.

В матрице $\{a_{ij}\}$ берется строка, соответствующая первой градации $i_1$ интересующего нас признака $l$ , т.е. выделяется разрез социологического массива по $i_1$ . В этой строке всего $n$ элементов-чисел $a_{ij}$ ( $j = 1,2, \ldots, n$ ), которые сгруппированы в $k$ групп , подобно тому, как все $n$ градаций анкеты сгруппированы в $k$ признаков. В каждой из этих $k$ групп чисел $i_1$ -ой строки выделяется максимальное число, затем отмечаются соответ-

ствующие градации; тем самым образуется множество из $k$ градаций. Это и есть искомое множество информативных градаций-характеристик для исходного таксона $i_1 \in \ell$. Обозначим его через $X_{i_1}$. Далее аналогичным образом определяются множества $X_{i_2}, \ldots, X_{i_{p_\ell}}$ наиболее информативных характеристик-градаций соответственно для остальных таксонов $i_2, i_3, \ldots, i_{p_\ell}$. После этого рассматривается пересечение полученных множеств $X_{i_1}, X_{i_2}, \ldots X_{i_{p_\ell}}$. Если оно не пусто, т.е. если $A = \bigcap_{i \in \ell} X_i \neq \Phi$, то $A$ есть искомое множество общих характеристик-градаций для признака $\ell$. Таким образом, из всех градаций, имеющихся в анкете, лишь градации, принадлежащие множеству $A$, наиболее типичны, например, для представителей всех национальностей. Если же $A = \Phi$, то можно предположить, что во множестве всех градаций анкеты нет таких градаций- характеристик, которые были бы в общем типичны для представителей каждой национальности.

Специфические же характеристики для каждой градации признака $\ell$ определяются следующим образом. Находятся объединения $B = \bigcup_{i \in \ell} X_i$, $C_1 = X_{i_2} \cup X_{i_3} \cup \ldots \cup X_{i_{p_\ell}}$, $C_2 = X_{i_1} \cup X_{i_3} \cup X_{i_4} \cup \ldots \cup X_{i_{p_\ell}}$ и т.д., наконец, $C_{p_\ell} = X_{i_1} \cup X_{i_2} \cup \ldots \cup X_{i_{p_\ell - 1}}$ и рассматриваются дополнения $B \setminus C_1$, $B \setminus C_2$, ..., $B \setminus C_{p_\ell}$, которые образуют множества в общем наиболее специфичных характеристик градаций соответственно для таксонов $i_1, i_2, \ldots, i_{p_\ell}$ признака $\ell$.

Таков алгоритм решения поставленной задачи. Он наиболее прост и потому "стоимость затрат" на его реализацию минимальна. Однако этот алгоритм обладает рядом существенных недостатков. Прежде всего следует отметить, что в данном алгоритме совершенно не учитывается возможная коррелированность признаков, а, следовательно, и возможная коррелированность между градациями различных признаков. В описанном случае алгоритмом выделяются лишь такие градации, которые информативны сами по себе, а не в отношении с другими градациями; тем самым производится

искусственное стирание всех связей между градациями различных признаков. Большим недостатком этого алгоритма является и то, что при выделении из признака наиболее типичной для исходного таксона характеристики-градации не учитывается количество градаций в этом признаке. Поэтому изменение числа градаций в каком-либо вопросе анкеты может повлечь за собой замену одной характеристики в искомом информативном множестве другой. Существенной погрешностью этого алгоритма является отсутствие явно выраженной зависимости искомых результатов от объема выборки. Однако, несмотря на наличие таких недостатков, этот алгоритм иногда целесообразно использовать для получения либо предварительных, схематичных результатов, либо - результатов, сравниваемых с аналогичными результатами, полученными другими методами и способами.

Описанный алгоритм является далеко не единственным способом решения сформулированной задачи. Мы остановимся на таком способе решения, в котором отсутствуют указанные недостатки.

Данные матрицы $\{a_{ij}\}$ позволяют различными способами ([5], стр. 59) определить показатели связи между градациями $i$ и $j$, которые можно интерпретировать как "расстояния" между $i$ и $j$. В качестве одного из таких показателей рассмотрим сначала простейший - разность между единицей и средней долей $a_{ij}$ в $a_{ii}$ и $a_{jj}$, т.е.

$$r(i,j) = \frac{a_{ij} - \frac{a_{ii} \cdot a_{jj}}{N}}{\left| a_{ij} - \frac{a_{ii} \cdot a_{jj}}{N} \right|} \left[ 1 - \frac{a_{ij}}{2} \left( \frac{1}{a_{ii}} + \frac{1}{a_{jj}} \right) \right] \qquad (I)$$

Здесь коэффициент $\frac{a_{ij} - \frac{a_{ii} \cdot a_{jj}}{N}}{\left| a_{ij} - \frac{a_{ii} \cdot a_{jj}}{N} \right|}$ определяет знак + или знак - для $r(i,j)$. Первый из них указывает на взаимосвязь градаций $i$ и $j$, а второй - на ковзаимосвязь, т.е. на разъединенность, разобщенность $i$ и $j$ (5). В дальнейшем для определенности полагаем $r(i,j) \geq 0$.

В выражении (1) произвольные, но фиксированные градации $i$ и $j$ рассматриваются не изолированно друг от друга, не сами по себе, а взаимосвязанно, взаимозависимо. Однако в формуле (I) не отражается зависимость информативности градаций $i$ и $j$ от объема выборки $N$, т.е. показатели близости $r(i,j)$ градаций $i$ и $j$ могут быть одинаковыми при опросе как малого количества респондентов, так и большого. Тем самым репрезентативность выборки никоим образом не учитывается в соотношении (I). Этот недостаток избегается небольшим усложнением формулы (I).

Общеизвестно, что при увеличении объема выборки $N$ возрастает репрезентативность по тому или иному признаку, а значит, и по каждой градации этого признака. Однако замечено, что при увеличении малых объемов выборки она возрастает больше, чем при увеличении больших объемов выборки, т.е. определенное числовое изменение малого объема выборки влияет на репрезентативность сильнее, чем то же самое изменение большой выборки. В практической социологии эта эмпирически устанавливаемая закономерность достаточно хорошо описывается логарифмической функцией $y = \log_N a_i$; где $y$ – показатель репрезентативности по признаку ( или) градации) $i$, $N$ – объем выборки, $a_i \leq N$ – количество ответов на признак (или градацию) $i$ при опросе $N$ респондентов. Очевидно, что $0 \leq y \leq 1$.

Если учесть логарифмическую зависимость репрезентативности по той или иной градации ( и даже по всем градациям!) то выражение (I) можер представить в виде:

$$r(i,j) = \pm\left[1 - \frac{a_{ij}}{2}\left(\frac{\log_N a_{ii}}{a_{ii}} + \frac{\log_N a_{jj}}{a_{jj}}\right)\right] = \pm\left[1 - \frac{a_{ij}}{2\ln N}\left(\frac{\ln a_{ii}}{a_{ii}} + \frac{\ln a_{jj}}{a_{jj}}\right)\right]. \quad (2)$$

В качестве показателя близости между градациями $i$ и $j$ берется разность между I и средней долей $a_{ij}$ в $a_{ii}$ и $a_{jj}$, логарифмически соотнесенных к объему выборки $N$.

Формула (2) достаточно полно описывает парные связи между градациями, но в (2) не рассматривается влияние количест-

ва градаций в признаках $l(i)$ и $l(j)$ на числовое выражение этих связей. Поэтому вместо формулы (2) целесообразно брать выражение:

$$r(i,j)=\pm\left[1-\frac{a_{ij}}{2\ln N}\left(\frac{\ln(p_{l(i)}\cdot a_{ii})}{a_{ii}}+\frac{\ln(p_{l(j)}\cdot a_{jj})}{a_{jj}}\right)\right]. \qquad (3)$$

Влияние количества градаций признака на показатель $r(i,j)$ иногда может и отсутствовать. Действительно, положим, что $a_{ii}=a_{ij}=a_{jj}$, тогда, очевидно, что градации $i$ и $j$ совпадают. Поэтому хотя бы в одном из признаков $l(i)$ или $l(j)$ градацию $i$ или $j$ можно опустить. Но так как в признаке все градации взаимоисключающиеся, то при "отбрасывании" одной градации могут измениться числа ответов на оставшиеся градации этого признака. Следовательно, "отбрасывание" одной из совпадающих градаций изменяет показатель $r(i,j)$ для оставшихся градаций этого же признака. Однако очевидно, что при совпадении $i$ с $j$ изменение количества градаций в $l(i)$ или $l(j)$ не вызовет изменения показателя близости $r(i,j)$ между градациями $i$ и $j$. . Поэтому, если $i$ совпадает с $j$, то в формуле (3) можно положить $p_{l(i)}=p_{l(j)}=1$. Вследствие этого выражение (3) для $a_{ii}=a_{ij}=a_{jj}$ переводится в соотношение (2).

Легко видеть, что в формулах (I), (2) и (3) $0\le|r(i,j)|\le 1$.

Таким образом, определяемый по формуле (I) показатель $r(i,j)$ равен $0$ тогда и только тогда, когда $a_{ii}=a_{ij}=a_{jj}$. В этом случае независимо от того, близки $a_{ii}$, $a_{ij}$ и $a_{jj}$ к объему выборки $N$ или очень малы в сравнении с $N$, показатель связи $r(i,j)=0$ как для малых чисел $a_{ii}=a_{ij}=a_{jj}$, так и для чисел, близких к $N$. Такой недостаток отсутствует в определении $r(i,j)$ по формуле (2); здесь $r(i,j)=0$ тогда и только тогда, когда $a_{ii}=a_{ij}=a_{jj}=N$. Так как формула (3) применима для $i$ не совпадающих с $j$, то в этом случае $r(i,j)>0$ всегда. В формулах (I), (2) и (3) $r(i,j)=1$, если $a_{ij}=0$. Так, например, если градации $i$ и $i+1$ выражают признак "пол" ( $i$ -"мужской", $i+1$ -

"женский"), то $a_{i\,i+1} = 0$, хотя $a_{ii} \neq 0$ и $a_{i+1\,i+1} \neq 0$. Следовательно, $r(i, i+1) = 1$, что означает отсутствие близости между градациями "мужской" и "женский" в признаке "пол".

Определив по одной из формул (1), (2) и (3) показатели $r(i,j)$, можно социологический массив представить в виде полного, симметричного, неориентированного, связного графа $G$ [3], [4]; где все градации анкеты образуют множество вершин, а иножество ребер образовано парными связями между градациями, определяемыми в любом из указанных соотношений. Введем разбиение графа $G$ по отношению инцидентности ребер вершинами из $l$ [3], [4]; обозначим получающиеся суграфы через $G_l$, где $l = 1, 2, \ldots, h$. Аналогично разобьем граф $G$ по отношению инцидентности ребер вершине $i$ и получающиеся суграфы обозначим через $G_i$, где $i = 1,2, \ldots, n$. Так как граф $G$ связен, то $G_i \cap G_l \neq \bar{\Phi}$. Это пересечение, образующее подграф, обозначим через $G_{il}$. Множество исходных характеристик $S$ отождествим со множеством всех вершин графа $G$.

Тогда решение поставленной задачи сводится к следующему.

В графе $G$ при заданных начальных вершинах $i_1$ (таксон "русские"), $i_2$ (таксон "киргизы"), $i_3$ (таксон "эстонцы") и т.д. $i_{p_1}$ (таксон "другие национальности") выделяем цепи, соответственно характеризующие эти вершины - таксоны. Общая часть таких цепей, если она имеется, будет раскрывать общие черты, присущие всем заданным таксонам, а те части, которые имеются лишь в одной из этих цепей и не имеются в остальных, характеризуют и выделяют специфические черты соответствующих таксонов, которые заданы. Эта задача решается алгоритмически.

В описании алгоритма условимся порядок отмечаемых вершин-градаций искомой цепи, а также порядок признаков, которым принадлежат эти градации, указывать индексацией букв $l$ и $i$.

В интересующем нас признаке $l$ выбираем произвольную градацию $i$, например, в признаке "национальность" гра-

дацию "другие национальности". Переобозначим их, т.е. для $l$ и для $i$ поставим индекс $1$, иначе - начальный признак будет $l_1$, а начальная градация - $i_1$. Тогда алгоритм кратко можно описать так.

В суграфе $G_{i_1}$ выбирается такое инцидентное вершине $i_1$ ребро, которое удовлетворяет требованиям:

а)соответствующее число $r(i_1, i_2)$, определяемое по одной из формул (I), (2) и (3), наиболее близко к нулю;

б) инцидентная этому ребру вершина $i_2 \in l_2$, но $i_2 \notin l_1$.

Замечание. При этом, если вершина $i_2$ есть такая градация, после которой в анкете следует вопрос (или группа их) вида "если не $i_2$, то почему" с определенным количеством градаций, то ребра, инцидентные этим вершинам-градациям, далее не рассматриваются. И обратно, если вершина $i_2$ есть градация признака вида "если $i_3$, то почему", то на следующем шаге алгоритмического процесса делается переход к рассмотрению вершины $i_3$. Для этого в программе, реализующей излагаемый алгоритм, достаточно предусмотреть соответствующие "переходы".

Такая вершина $i_2$ отмечается в подграфе $G_{i_1 l_2}$, затем в суграфе $G_{i_2}$ аналогичным образом выбирается ребро, удовлетворяющее и условию а), и условию б) - $i_3 \in l_3$, но $i_3 \notin l_1$, $i_3 \notin l_2$. После чего вершина $i_3$ отмечается в подграфе $G_{i_2 l_3}$. Этот процесс продолжаем до тех пор, пока не отметится последняя вершина $i_k \in l_k$, не принадлежащая $l_1, l_2, \ldots, l_{k-1}$. Для полученной цепи (обозначим ее $Simp_1$ ( $i_1, i_2, \ldots, i_k$ ) с $k$ отмеченными вершинами определяется число $t_1 = \sum_{\alpha=1}^{k-1} r(i_\alpha, i_{\alpha+1})$

Затем в суграфе $G_{i_1}$ берется ребро, у которого $r(i,j)$ - второе из наиболее близких к нулю чисел, и дальше совершенно идентичным способом определяется вторая цепь $Simp_2$ с числом $t_2$. Такой процесс выделения цепей в графе $G$ при заданной начальной вершине $i_1$ продолжается до перебора всех $n-p_{l_1}$ ребер в суграфе $G_{i_1}$. Цепь $Simp$ из цепей $Simp_1, Simp_2, \ldots, Simp_{n-p_{l_1}}$ с числом $t = \min\{t_1, t_2, \ldots, t_{n-p_{l_1}}\}$ будет искомой цепью, которая ха-

рактеризует заданный начальный таксон ( в нашем примере таксон "другие национальности").

Аналогичным образом определяются остальные $p_\ell - 1$ цепи, характеризующие соответствующие заданные начальные градации-таксоны признака $\ell$ ("национальность").

Таким образом, для определения $p_\ell$ информативных цепей требуется выделить $p_\ell (n - p_\ell)$ цепей длины $k$ в графе $G$ почти из $n!$ возможных. Так как число $n$ обычно бывает порядка нескольких сотен, а $p_\ell$ редко превышает число 10, то величина $p_\ell (n - p_\ell)$ очень мала в сравнении с $n!$, а поэтому и "стоимость затрат" $R$ на реализацию алгоритма выделения информативных цепей относительно низкая.

Рассмотрим пересечение $p_\ell$ полученных в графе $G$ цепей. Если оно существует, то такое пересечение выражает общие черты, характерные для всех таксонов в признаке $\ell$, в противном случае можно предположить, что в анкете нет таких характеристик, которые были бы общими для всех таксонов признака $\ell$. Если в каждой искомой цепи имеются такие части, которые не являются частями всех других искомых цепей, то эти части выражают специфические черты таксонов, соответствующих градациям признака $\ell$.

Таково краткое описание одного из способов решения задачи о выделении информативного подмножества характеристик $X \subset S$. Однако здесь нами не решается важная для практической работы социолога задача оценки информативности подмножества $X$. Ввиду сложности определения оценки информативности подмножества $X \subset S$ такая задача требует специального рассмотрения.

Существенным достоинством изложенного способа решения социологических задач подобного типа является то, что в данном случае одновременно могут рассматриваться и анализироваться признаки как с метрическими, так и ранжированными и даже классификационными типами шкал [6. стр. 18 - 27], [7. стр. 10 - 12]. Однако

этот способ решения задач такого типа недостаточно опробован нами в практике обработки социологической информации, хотя первые полученные результаты согласуются и с другими результатами и с экспертно-интуитивными представлениями.

ЛИТЕРАТУРА


I. Н.Г. Загоруйко. Классификация задач распознавания образов. - Вычислительные системы, Новосибирск, I966, вып. 22.

2. Н.Г. Загоруйко. Современное состояние проблемы распознавания образов. - Вычислительные системы, Новосибирск, I967, вып. 28.

3. О.Оре . Теория графов, М., I968.

4. А.А. Зыков. Теория конечных графов, Новосибирск,I969.

5. Д.Юл, М. Кендэл. Теория статистики,М.,I960.

6. П. Суппес, Дж. Зинес. Основы теории измерений. - В сб.: Психологические измерения,М.,I967.

7. Ю.П. Воронов, Н.П. Ершова. Общие принципы социологического измерения. В сб.: Измерение и моделирование в социологии, Новосибирск,I969.

## INFORMATIIVSETE KARAKTERISTIKUTE ERALDAMISEST SOTSIOLOOGIAS

A. I. Tišin
(Frunze)

Resümee

Karakteristikute informatiivse alamhulga väljaselgitamise ülesanne esineb sotsioloogias küllaltki sageli.

Käesolevas artiklis esitatakse üks võimalik algoritm selle ülesande lahendamiseks. Aluseks võetakse n x n sagedusmaatriks, kus

$$n = \sum_{l=1}^{\kappa} p_l,$$

$p_l$ on $l$-nda tunnuse väärtuste (gradatsioonide) arv, $l = 1,2,\ldots,\kappa$, $\kappa$ on tunnuste arv. Selle maatriksi ridade analüüsimisel leitakse iga tunnuse iga gradatsiooni jaoks iga teise tunnuse informatiivseim väärtus.

Edasi varieeritakse toodud algoritmi eesmärgiga vähendada tulemuste sõltuvust väljavõtte mahust ja üksiktunnuste väärtuste arvust.

## THE DISTINCTION OF INFORMATIVE CHARACTERISTICS IN SOCIOLOGY

A.I. Tishin
(Frunze)

Summary

The problem of finding the informative subset of characteristics arises often in sociology.

One possible algorithm for solving this problem is given in this article. The algorithm is based on the $n \times n$ frequency matrix where

$$n = \sum_{l=1}^{\kappa} p_l$$

$p_l$ is the number of gradations of $l$-th variable.

By analysing the lines of this matrix for each gradation of a variable the most informative gradation of another variable is found.

Then the given algorithm is variated with the aim of diminishing the dependence of results on the size of the sample and the number of gradations.

# О СРАВНЕНИИ РАЗЛИЧНЫХ МЕТОДОВ КЛАССИФИКАЦИИ В СОЦИОЛОГИИ

(в порядке постановки и обсуждения)

М.Р. Ланин (Тарту), А.И. Тишин (Фрунзе)

Идея совместной работы по проверке эффективности различных методов классификации возникла в процессе многочисленных дискуссий, имевших место летом 1970 года в Новосибирской школе-семинаре "Социология и математика".

Во многих кибернетических, вычислительных и социологических центрах страны и за рубежом накопилось некоторое количество алгоритмических методов и программ классификации. Однако применение их к обработке социологического материала обычно дает различный эффект, что затрудняет получение правильной, объективной социологической интерпретации обрабатываемой информации. В свою очередь имеется и достаточное число показателей связи между различными признаками на одном и том же социологическом материале, т.е. на одном и том же множестве социологических признаков вводятся самые разнообразные "метрики"[I], но вопросы о сравнении и тем более согласованности их остаются до сих пор открытыми. В связи с этим интересно выяснить: какие методы классификации следует использовать при выборе различных "метрик", а также - возможны ли, и если возможны, то каковы иерархии "метрик", методов классификации и "метрик-методов". Иными словами, как следует упорядочить "метрики" и методы классификации, чтобы найти предпочтительный метод классификации для заданной "метрики" и обратно - предпочтительную "метрику" для данного метода классификации.

---

I В определении "метрик" на множестве социологических признаков очень часто отсутствует аксиома транзитивности.

Пытаясь экспериментально решить эти задачи, мы начали обрабатывать на ЭВМ большой социологический материал, полученный в результате исследований национальных отношений в Киргизии. При этом делается следующее: вычисляется несколькими способами количественное выражение связи между различными признаками, оформленными в виде вопросов анкеты, т.е. на одном и том же множестве социологических признаков, вводятся различные "метрики".

В качестве показателей таких связей или "метрик" берутся:

I. Парные коэффициенты корреляции, определяемые по следующей формуле:

$$\eta_{x,y} = \frac{\sum_{i=1}^{p}\sum_{j=1}^{s}(x_i-\bar{x})(y_j-\bar{y})\,b_{ij}}{\sqrt{\sum_{i=1}^{p}(x_i-\bar{x})^2 b_i \sum_{j=1}^{s}(y_j-\bar{y})^2 b_j}} \qquad (1)$$

где $(x_1, x_2, \ldots, x_p)$ – вектор-признак $X$;

$(y_1, y_2, \ldots, y_s)$ – вектор-признак $Y$;

$$\bar{x} = \frac{\sum_{i=1}^{p} x_i}{p}; \qquad \bar{y} = \frac{\sum_{j=1}^{s} y_j}{s};$$

прямоугольная матрица $\{b_{ij}\}$ представляет собой корреляционную таблицу, образованную признаками $X$ и $Y$. В данном случае коэффициенты $\eta_{x,y}$ показывают и на тесноту связи между качественными признаками.

II. Коэффициент связи вида[2].

$$\rho(X,Y) = \frac{1}{R(R-1)}\left[\sum_{j=1}^{s}\left(\sum_{i=1}^{p} b_{ij}\right)^2 + \sum_{i=1}^{p}\left(\sum_{j=1}^{s} b_{ij}\right)^2 - 2\sum_{i=1}^{p}\sum_{j=1}^{s} b_{ij}\right] \qquad (2)$$

где $R$ – объем выборки.

Отличительной чертой показателя $\rho(X,Y)$ является простота его вычисления, кроме того, для $\rho(X,Y)$ также безразлично, являются ли $X, Y$ количественными, ранжированными или качественными признаками.

---

[2] Б.Г.Миркин. Об одном подходе к анализу первичной социологической информации. Новосибирск, 1970.

III. Корреляционное отношение:

$$\eta = \sqrt{\frac{\delta^2}{\sigma^2}},$$

где: $\bar{\delta}^2$ - межгрупповая дисперсия, $\delta^2 = \frac{\sum_{i=1}^{s}(\bar{y}_i - \bar{y})b_i}{\sum_{i=1}^{p}\sum_{j=1}^{s} b_{ij}}$ (3);

$\sigma^2$ - общая дисперсия, $\sigma^2 = \bar{\delta}^2 - \bar{\sigma}^2$;

$$\bar{\sigma}^2 = \frac{\sum_{i=1}^{s} \sigma_i^2 b_i}{\sum_{i=1}^{p}\sum_{j=1}^{s} b_{ij}}.$$

Формулы (I), (2) и (3) дают показатель связи между каждой парой признаков. Если в анкете $n$ признаков, то все такие связи можно представить в виде трех квадратных матриц порядка $n$. К каждой из полученных матриц применимы следующие методы классификации:

А. "Максимум корреляционного пути"[3], сущность которого сводится к следующему. Связываем все $n$ признаков системы при помощи $n-1$ связей так, чтобы сумма коэффициентов корреляции связанных между собой признаков была максимальной.

В. Метод факторного анализа ($Q$ - техника).[4]

С. Алгоритм классификации объектов с помощью методов теории графов, изложение которого печатается в данном сборнике (см. стр. ).

Таким образом, исходя из трех способов задания "метрики" и трех методов классификации, мы получаем девять видов вообще говоря не совпадающих результатов:

| | | |
|---|---|---|
| IA | IB | IC |
| IIA | IIB | IIC |
| IIIA | IIIB | IIIC |

(4)

---

[3] Л. Выханду. Об исследовании многопризнаковых биологических систем. В сб.: Применение математических методов в биологии, т. 3. Изд. ЛГУ, 1964.

[4] Лоули Д.Н. и Максвелл А.Э. Факторный анализ как статистический метод. М., "Мир", 1967.

Количество этих результатов можно значительно расширить, рассмотрев дополнительно еще какую-либо "метрику", например, коэффициент корреляции по Чупрову или - какой-либо метод факторного анализа.

Тогда задача сводится, во-первых, к выделению в таблице (4) результата, наиболее адекватно отражающего исследуемый социальный процесс или явление, и, во-вторых, к выделению аналогичных результатов в каждой строке и каждом столбце этой же таблицы. Такая задача может быть решена как формальными, так и содержательными методами. Наиболее простыми из них являются методы экспертных оценок и интуитивных представлений.

В заключение отметим, что работа в этом направлении продолжается и окончательные выводы могут быть сделаны лишь после проверки излагаемой методики на различных социологических массивах.

## MÕNEDE KLASSIFITSEERIMISMEETODITE VÕRDLEMISEST SOTSIOLOOGIAS

M.R. Lanin (Tartu),
A.I. Tišin (Frunze)

Resümee

Töös püstitatakse ülesanne konkreetsel sotsioloogilisel materjalil kontrollida mõnede klassifitseerimismeetodite efektiivsust. Kasutatakse mitmeid erinevaid tunnustevahelise seose karakteristikuid.

## THE COMPARISON OF SOME CLASSIFICATION METHODS IN SOCIOLOGY

M.R. Lanin (Tartu),
A.J. Tishin (Frunze)

Summary

In the paper the effectiveness of some classification methods is proved on concrete sociological material. Several characteristics of interrelations between the variables are considered.

# ВЫДЕЛЕНИЕ СОЦИАЛЬНО-ДЕМОГРАФИЧЕСКИХ ТИПОВ МЕТОДАМИ КЛАСТЕР-АНАЛИЗА И ОПРЕДЕЛЕНИЕ ИХ СВЯЗИ С ТИПАМИ ПОВЕДЕНИЯ

Л. Гордон, Э.Клопов, А. Терехин,
М. Скворцов

( Москва)

I. Разработка проблем социальной структуры зрелого социалистического общества составляет одно из центральных направлений советской социологии. Весьма важное значение в связи с этой проблемой имеет "вычленение" основных элементов социальной структуры, разбиение всей массы членов общества на группы, состоящие из людей с более или менее сходными социальными характеристиками. Выделение таких групп создает предпосылки для установления связей между ними и, следовательно, для анализа социальной структуры в собственном смысле слова.

2. Однако "объединение" людей в сравнительно однородные социальные группы требует преодоления весьма серьезных трудностей. Едва ли не самая существенная в их ряду проистекает из множественности значимых социально-демографических характеристик, которые должны учитываться при выделении элементов социальной структуры - относительно однородных социальных групп. Люди, как объекты социально-демографической классификации, выступают в виде существенно многомерных образований.

2.I. Необходимость учета многих социально-демографических показателей с особой силой проявляется при определении элементов социальной структуры современного советского общества. В условиях развитого классового общества - например, капиталистического или даже переходного к социализму - положение существенно упрощается, поскольку здесь теоретически известна некоторая иерархия значимости социальных характеристик. Доказано, что в подобных обществах классовая принадлежность, а также принадлежность

к основным внутриклассовым слоям имеет гораздо большее значение для определения места человека в социальной структуре, чем прочие социально-демографические признаки. Подобная иерархированность дает возможность сначала разделить всех членов общества на классы, в зависимости от показателей, характеризующих классовую принадлежность; затем выделить основные внутриклассовые слои и т.д. При этом каждый шаг такой классификации требует учета не всего множества, но лишь одного-двух показателей.

Однако в условиях классово-однородных социальных общностей прежняя иерархия неприменима. При выделении социальных групп в социалистическом городе нельзя говорить о решающем значении признака классовой принадлежности, поскольку этот признак одинаков у всех или почти у всех городских трудящихся. Что касается остальных явно значимых социально-демографических характеристик, их иерархия, их относительная роль в формировании элементов социальной структуры в настоящее время далеко не очевидна. Пока мы не имеем оснований утверждать, что в этом отношении квалификация,например,всегда важнее образования или наоборот. "Вычленение" относительно однородных социальных групп в социалистическом городе требует принимать во внимание именно неиерархированное множество социально-демографических характеристик (собственно, лишь после такой операции мы можем надеяться на получение убедительных оснований для введения иерархии этих признаков).

2.2. Необходимость учета весьма большого числа показателей (существенная многомерность объектов социальной классификации) закрывает еще один путь, который мог бы облегчить дело, если бы речь шла о небольшом числе характеристик. В этом последнем случае можно было бы свести классификацию к разбиению людей на абсолютно однородные в социальном смысле группы, т.е. группы, у которых все характеристики совершенно одинаковы. ( Скажем, группа работников средней квалификации, занятых механизированным трудом, имеющих образование 7 - 9 классов, в возрасте 35 - 45 лет, имеющих семью и

несовершеннолетних детей, живущих в крупных городах, имеющих душевой доход в семье от 50 до 75 руб. в месяц и т.д. и т.п. по всем измеряемым социально-демографическим характеристикам).

Однако при мало-мальски значительном числе характеристик такой подход становится практически невозможным вследствие чрезвычайной многочисленности абсолютно-однородных групп. При 40 - 50 характеристиках типа начальное, среднее, высшее образование, низшая, средняя, высокая квалификация и т.п., общее число возможных комбинаций составит $2^{40} - 2^{50}$.

Число же социально-демографических характеристик не может быть значительно уменьшено, во-первых, из-за неиерархированности признаков и вытекающей отсюда невозможности "выбросить" неважные показатели и, во-вторых, из-за неочевидности принципов укрупнения родственных признаков, путей их "склеивания" друг с другом (неясно, например, следует ли начальное образование объединять со средним или среднее с высшим, среднее образование со средней квалификацией или высокой и т.п.).

Короче, и здесь мы приходим к необходимости учитывать возможно большее число социально-демографических характеристик.

3. Итак, природа объектов социальной классификации сводит проблему выделения элементов, составных частей социальной структуры ( по крайней мере, при нынешнем уровне развития теории) к поиску естественного, реального набора относительно однородных социальных групп, состоящих из людей с более или менее близкими, сходными характеристиками, на которые распадаются люди ( носители этих характеристик) в реальной действительности. Задачи подобного рода обычно решаются методами многомерной статистики, в частности методами распознавания образов, такими как кластер-анализ, таксономия, автоматическая классификация (здесь и далее эти термины употребляются как равнозначные)[1].

[1] Строго говоря, эти методы относятся не столько к распознаванию образов, сколько к формированию, выделению образов, так сказать, к задачам, обратным распознаванию образов. Однако первое обозначение получило уже всеобщее хождение в реальном словоупотреблении.

3.I. Казалось бы, простейший путь решения нашей задачи должен состоять в прямом таксономическом разбиении социально-демографических данных исследуемой совокупности. Каждого человека в этом случае следовало бы рассматривать как своего рода точку, вектор с координатами, соответствующими значениям всех его социальных, демографических и т.п. характеристик, учитываемых нами. Расположение подобных многомерных точек друг относительно друга в принципе позволяет определить их естественные скопления. Разумно считать, что разделение точек на скопления отражает разделение людей, которые обозначаются этими точками, на естественные группы. В каждую такую группу входят точки с близкими координатами и соответственно люди с более или менее близкими, сходными социально-демографическими характеристиками. В этом смысле здесь и появляются искомые относительно однородные группы, образующие элементы, "слагаемые" социальной структуры.

3.2. Однако подобный путь несмотря на его кажущуюся простоту и очевидность приводит к почти непреодолимым препятствиям. Расположение многомерных векторов в пространстве их свойств существенно зависит от того, считаем ли мы те или иные социальные характеристики (образование, квалификацию, материальное положение и т.п.) равнозначащими при определении близости, похожести людей друг на друга или нет. Из общих соображений и здравого смысла, ясно что значение большинства социально-демографических показателей неодинаково в процессе разбиения людей на относительно сходные группы. К сожалению, нет никаких очевидных оснований, чтобы определить количественную меру значимости того или другого показателя, его вес при определении относительного сходства.

Возникающая здесь опасность субъективных ошибок и произвола чрезвычайно велика. Утверждение о разнозначности всех показателей есть, в сущности, столь же произвольное решение, как и придание им разных весов. Соответственно и все разбиение точек на группы ( а значит и выделение однородных в социальном отношении группы) приобретает субъек-

тивно-произвольный характер, так как придание того или иного веса различным координатам определяет метрику пространства свойств, в котором расположены все эти точки.

3.3. В данной связи представляется, что одним из способов частичного преодоления трудностей метрики (точнее их "обхода") является изменение исходной задачи. Откажемся от прямой попытки разбиения людей на относительно однородные в социальном отношении группы. Вместо этого зададимся целью найти такие сочетания социально-демографических характеристик, которые чаще всего встречаются в исследуемой действительности. (Допустим, сочетание неполного среднего образования с высокой квалификацией, средним возрастом, семейностью, относительно высоким доходом, партийностью, хорошими жилищными условиями или сочетание низшего образования с низкой квалификацией, пожилым возрастом и т.д.). При таком подходе появляется определенное основание отвести вопрос об относительной значимости тех или иных характеристик. Ведь в этом случае мы не пытаемся определить относительную похожесть, близость людей, но стремимся определить относительно чаще встречающиеся социальные типы, выявить некоторые живые сочетания социальных характеристик, реально бытующие в действительности.

С известным огрублением можно сказать, что мы возвращаемся к выделению абсолютно однородных групп (где проблема весов и метрики снимается), но берем не все такие группы - что невозможно - но лишь некоторые из них, те, которые соответствуют чаще всего встречающимся сочетаниям социально-демографических свойств.

Разумеется, подобный обход главной трудности неизбежно ведет к значительным потерям. Задача разбиения всех людей (всех точек) на группы остается неременной, она попросту снимается. В случае успеха достигается определение лишь отдельных типов, в которых отражаются отнюдь не все реально существующие сочетания социальных свойств. (Например, если среднее образование чаще всего сочетается с высокой квалификацией, то достаточно часто, хотя и реже, оно сочетается с квалификацией низкой или средней). Более того, самые харак-

терные и чаще всего встречающиеся сочетания, как правило будут охватывать лишь меньшинство любой исследуемой совокупности; общее число возможных сочетаний чрезвычайно велико и потому невероятно, чтобы немногие наиболее характерные из них охватывали слишком большое число точек. Однако любое другое сочетание социальных свойств окажется менее характерным, любой другой тип будет встречаться в действительности еще реже.

Иными словами, можно считать, что сочетания социальных характеристик, чаще всего встречающиеся в действительности, дают представление о наиболее характерных социально-демографических типах, рисуют, своего рода, идеальные прототипы, образы идеальных представителей важнейших социально-демографических групп. Мы не получаем здесь всего перечня относительно однородных социально-демографических групп, но выявляем, так сказать, координаты, точки отсчета, между которыми эти группы расположены.

В этом смысле замена задачи разбиения совокупности людей на относительно однородные группы задачей выделения социально-демографических типов, соответствующих чаще всего встречающимся сочетанием социальных свойств, означает важный шаг по пути эмпирического определения элементов социальной структуры и их связи друг с другом.

3.4. (Во избежание недоразумений заметим, в скобках, что предлагаемая в пункте 3.3. замена задач не есть единственная возможность преодоления "проклятия метрики", необходимости определить количественное соотношение значимости отдельных характеристик, о котором говорилось выше).

4. Задача поиска таких сочетаний социально-демографических свойств, которые относительно чаще встречаются в действительности, решается методом кластер-анализа, где в качестве кластеров (таксонов) выступают не точки со множеством характеристик, но определенные сочетания этих характеристик, как они отражаются корреляционными отношениями.

4.1. Конкретно, решение достигается следующим образом. Рассматривается матрица парных корреляций всего множества учтенных социально-демографических характеристик исследуе-

мой совокупности (т.е. фактически сведения о том, какие из всех возможных парных сочетаний чаще и какие реже встречаются в действительности). В результате рассмотрения на первом шаге выделяется пара характеристик с наиболее высоким коэффициентом корреляции. (Например, возраст до 25 лет и членство в ВЛКСМ). Наибольшая величина данного коэффициента корреляции означает, что именно эти характеристики чаще всего сочетаются в исследуемой совокупности. Далее рассматривается новая матрица корреляций, в которой вместо двух характеристик с наиболее высоким коэффициентом корреляции действует одна объединенная характеристика ( в нашем примере, так сказать, "молодость-комсомольство").

В новой матрице повторяется прежняя процедура и снова выделяются чаще всего сочетающиеся характеристики. Это могут быть как сочетания первичных характеристик, так и сочетания вновь образованной характеристики ("молодость-комсомольство") с одной из первичных.

Операция повторяется до исчерпания (полного перебора) всех характеристик, то есть до полного слияния их в одно объединение, характеризующее исследуемую совокупность как целое[2].

4.2. Полученные результаты удобно представить в виде особой схемы - дерева, ветви которого показывают, какие признаки и совокупности признаков чаще сочетаются друг с другом.

Содержательно-социологический анализ подобного дерева облегчает выделение таких сочетаний социально-демографических характеристик, которые можно рассматривать как наборы, выявляющие относительно однородные социальные типы; вместе с тем дерево показывает взаимоотношение этих типов друг с другом.

4.3. Следует признать, однако, что здесь возникает известная трудность. На усмотрение исследователя остается

---

[2] Алгоритм, подобный изложенному п. 4.I., был впервые описан в "Scientific bull.of the Univ. of Kansas",1958,vol. 38,p. 1409 - 1438.

вопрос, какой уровень дерева, какой шаг объединения характеристик считать наиболее подходящим для выделения относительно однородного набора. Или говоря иначе: можно ли считать достаточным основанием для выделения социального типа сочетания первого уровня (например, "молодость-комсомольство") или таким основанием должно быть более обширное объединение ("молодость-комсомольство-хорошая обрзованность") - или еще более широкое сочетание ("молодость-комсомольство-хорошее образование-совмещение труда с учебой-средняя квалификация") и и т.д.

Ответ на этот вопрос принципиально лежит вне сферы формального анализа матрицы корреляций. Он должен базироваться на содержательных соображениях, учете задач классификации и т.д. В частности, для наших целей важно помнить о том, что в анализе участвуют характеристики, входящие в несколько основных категорий - есть показатели, дающие представление о квалификации (3 показателя), степени механизации труда (2 показателя), образовании (4 показателя) и т.д. Целесообразно выделить те сочетания, те уровни дерева, в которых появляются характеристики, относящиеся к каждой из этих основных категорий. Ибо только при этом условии мы можем рассматривать сочетание социально-демографических свойств как целостное отражение определенного социально-демографического типа, в котором учтены все важнейшие социальные характеристики. В противном случае, речь должна идти так сказать, о частичных, "усеченных" типах (отражена, например, квалификация и образование, но нет данных о семейном положении и др.).

5. Метод, описанный в п.4.I. и 4.2. был проверен в ходе анализа социально-демографических характеристик 300 работников - мужчин 9 крупных промышленных предприятий гг. Москвы, Днепропетровска, Запорожья, Одессы, Костромы, Павловского Посада.

5.I. Исследуемая совокупность образована в ходе вторичного отбора среди более чем 2000 рабочих, ИТР и служащих упомянутых предприятий. С вероятностью P = 0,9 можно

утверждать, что вторичная выборка по своим социально-демографическим характеристикам отличается от всей совокупности работников производственных цехов, обследованных предприятий, не более, чем на 5%.

В ходе опроса учитывалось 42 социально-демографические характеристики; подробно они перечислены в пояснениях к рис. I.

5.2. В итоге кластер-анализа указанных характеристик было получено дерево корреляций, показывающее, какие из 42 учтенных показателей (социальных свойств) чаще и какие реже сочетаются друг с другом ( см. рис. I). Рассмотрение дерева позволяет легко выделить три основных набора, сочетающихся друг с другом социальных характеристик (три главных группы ветвей дерева). В этих наборах представлены показатели всех основных категорий социальных признаков. Соответственно каждый из них можно рассматривать как отражение определенного социального типа, реально присутствующего в исследуемой совокупности. В свою очередь основные наборы, сочетания характеристик (основные ветви дерева) делятся на более мелкие (каждая из основных ветвей складывается из нескольких более мелких ветвей). Эти последние показывают как именно, в каком порядке и из каких составных частей (составляющих) формируются основные типы. В одних случаях такие составляющие характеризуют подтипы, в других - своего рода укрупненные социальные факторы, стоящие за данным блоком эмпирических показателей.

5.2.I. Первый из социально-демографических типов нашей совокупности характеризуют сочетания социальных свойств, сосредоточенные на центральных ветвях дерева.Они показывают, что к этому типу относятся люди со следующими показателями:

- высокая квалификация рабочих, высокие заработки (более I50 руб.), место жительства - большие города, высокая степень включенности в общественно-политическую деятельность (членство в КПСС и выполнение общественных поручений), проживание в отдельной квартире, пожилой возраст (свыше 45 лет) - первая составляющая данного типа;

- принадлежность к слою инженерно-технических работников и руководителей производства, высшее и среднее спецобразование - вторая составляющая;

- семейное положение, соответствующее центральному этапу жизненного цикла, наличие в семье одного несовершеннолетнего ребенка, средние заработки (IOI - I50 руб.) и средний уровень душевого дохода (5I - 75 руб.), наличие в семье телевизора - третья составляющая.

Таким образом, здесь налицо социально-демографический прототип наиболее развитой части промышленных рабочих и примыкающей к ним производственной интеллигенции; люди, которые находятся преимущественно на центральном этапе семейно-возрастного цикла, живут в условиях относительно высоко урбанизированной среды, сильнее других включены в современную городскую культуру социалистического общества.

5.2.2. Во втором типе (левые ветви дерева корреляций на рис. I) представлены рабочие, среди характеристик которых чаще всего встречаются следующие:

- принадлежность к группе рабочих ручного труда или вспомогательного труда у машин и механизмов, низкое образование (4 класса и ниже), проживание в общих коммунальных квартирах - первая составляющая;

- невысокая квалификация или отсутствие квалификации, низкие заработки (менее I00 руб.), слабоурбанизированная среда проживания (малый город) - вторая составляющая;

- обитание в пригородных зонах крупных городов, наличие собственного дома и подсобного хозяйства, образование 5 - 7 классов, беспартийность - третья составляющая.

Набор характеристик здесь не оставляет сомнений в определении места второго типа в социальной структуре: это сочетание, рисующее обобщенный образ менее развитой части рабочих, менее квалифицированных и соответственно хуже оплачиваемых, что вполне согласуется с их низким образованием; среди них много живущих в условиях, в сущности, полугородской среды.

5.2.3. Третий тип (правые ветви дерева) включает, во-первых, те социально-демографические показатели, которые

характерны, главным образом, для молодежного слоя рабочего класса: первая составляющая этого типа - молодость (здесь особенно много людей в возрасте до 26 лет) и отсутствие собственной семьи, членство в ВЛКСМ, среднее образование, совмещение труда с учебой, высокий душевой доход ( свыше 75 руб.) ,проживание в общежитиях.

А во-вторых, к нему же относятся поккзатели социально-профессионального статуса, характерные не только для молодежи, но среди молодых рабочих чаще встречающиеся: это - механизированный труд средней квалификации ( вторая составляющая).

5.2.4. Как видно из рис. I, не все первичные характеристики оказались включенными в три основных типа. Например, возраст от 26 до 35 лет не примыкает - до весьма высокого уровня - ни к одной из ветвей дерева. Следовательно, для рабочих этого поколения соотношение включенных в анализ социально-демографических и социально-профессиональных характеристик такое же, как и во всей обследованной совокупности.

5.3. Еще раз подчеркнем, что мы не разделили всех обследованных на группы, но выявили наиболее характерные, так сказать, идеальные социально-демографические типы. Конкретные социальные показатели большинства обследованных в тех или иных деталях отличаются от идеальных наборов. Однако эти последние есть своего рода центры, к которым тяготеет основная масса обследованных. Именно поэтому основные социально-демографические типы могут служить "точками отсчета", координатами для дальнейшей социальной классификации (т.е. для выделения собственно однородных социальных групп).

6. Помимо создания основ для эмпирического определения относительно однородных групп, выявление наиболее характерных сочетаний социально-демографических признаков оказывается весьма полезным при решении еще одной важной задачи - уяснении некоторых сторон связи между поведением и социальным положением, социальной почвой.

6.I. В свое время нами (совместно с В.Я. Волком и его сотрудниками) была предложена методика выявления групп, отличающихся более или менее одинаковым типом времяпрепровождения (разбиение бюджетов времени на сравнительно близкие группы с помощью методов распознавания образов). Методика эта была использована для анализа бюджетов времени ( в их внерабочей части) той же выборочной совокупности работников промышленных предприятий, о которой шла речь в п. 5. (Подробнее см. журнал "Вопросы философии" I969, № 7).

6.2. Результатом анализа явилось разбиение рассмотренных бюджетов времени на 5 категорий. Каждая из этих категорий может рассматриваться как особый тип поведения вне сферы производства, а люди с данным типом поведения - как специфическая группа поведения.

Первому типу использования внерабочего времени свойственна отчетливая семейно-домашняя направленность: в поведении людей этой группы большую роль играет семейное общение, особенно занятия с детьми. Для них характерны также очень значительные затраты времени, проводимого перед экраном телевизора.

Второй тип времяпрепровождения также имеет преимущественно "домашнюю" ориентацию. Однако она проявляется не столько в общении с детьми или "телефизионных" формах приобщения к культуре, сколько в относительно большом удельном весе домашних работ, особенно труда в подсобном хозяйстве. Поэтому в отличие от первого, "семейно-домашнего"типа, второй можно назвать "хозяйственно-домашним". Кроме того в этом случае значительно выше продолжительность и значение внесемейного общения.

Третий тип отличается от двух предыдущих сравнительно более развитой и разумно сбалансированной структурой затрат времени. Значительная интенсивность внесемейного общения сочетается с большими затратами времени на потребление культурных ценностей, в том числе на образование в системе вечернего и заочного обучения. В этом смысле здесь можно говорить о гармоническом типе времяпрепровождения.

Четвертый тип характерен для "рабочих-учащихся","рабочих-студентов". Это обстоятельство решительным образом сказывается на всем их внерабочем времени, придавая ему существенно "учебный" характер.

Наконец, пятый тип отличается значительной разнородностью времяпрепровождения. Тем не менее можно отметить, что для его структуры характерны многие черты, свойственные второму типу, однако, представлены они в более резкой форме. Это проявляется, в частности, в большой продолжительности труда в домашнем хозяйстве и особенно в чрезвычайной интенсивности внесемейного и внедомашнего общения.

7. Таким образом для одной и той же выборочной совокупности были установлены как наиболее характерные социально-демографические типы, так и основные типы поведения (времяпрепровождения). Подобное положение открывает возможность решить вопрос о наличии или отсутствии связи между теми и другими категориями. Представляется, что именно решение этого вопроса дает основание для более или менее строгого эмпирического подхода к проблеме зависимости поведения от социальной почвы, "жизненных обстоятельств ", стоящих за социально-демографическими характеристиками человека.

7.I. Правда, кажется, что зависимость поведения от условий жизненной обстановки может быть установлена и более простым способом - посредством анализа связи между каждой из социально-демографических характеристик и принадлежностью к группе поведения. Однако такой подход приводит к удовлетворительным результатам лишь в том случае, если эти связи чрезвычайно тесны, то есть если каждой группе (каждому типу) поведения соответствуют резко различные социально-демографические характеристики.

Фактически подобное положение встречается скорее как исключение, чем как правило, поскольку поведение непосредственно зависит не только от социально-демографической почвы, но и от характеристик сознания, культуры и других факторов, не полностью отражаемых в социально-демографических показателях.

7.2. В этих условиях решением вопроса является установление связи ( или ее отсутствие) между типами поведения и социально-демографическими типами, именно сочетаниями, блоками социальных характеристик, а не теми или иными показателями, взятыми в отдельности.

Для установления подобной связи достаточно повторить процедуру кластер-анализа матрицы корреляций, описанную в п. 4.I., включив однако в нее 5 дополнительных характеристик, обозначающих принадлежность или непринадлежность к той или иной группе поведения.

Такой анализ позволяет выяснить, можно ли утверждать, что принадлежность к определенной группе поведения чаще сочетается (или не сочетается) с определенным социально-демографическим типом. Если подобное положение обнаруживается, по нашему мнению, есть основание говорить о непосредственной социально-демографической почве, непосредственных "бытийных" предпосылках того или иного типа поведения.

7.3. Фактический анализ показал, что 3 из 5 групп поведения достаточно четко сочетаются с очерченными выше социально-демографическими типами (см. рис. 2).

Разумеется здесь нет абсолютного совпадения. Однако,как видно из центральной части рис. 2, принадлежность к первому типу поведения (семейно-домашняя ориентация) чаще чем в других случаях сочетается с такими свойствами, как средние размеры душевого дохода ( 5I-75 руб.), заработки - I0I - I50 руб., наличие семьи и одного ребенка, средний возраст,владение телевизором. Короче, принадлежность к этому типу поведения теснее всего связана с третьей составляющей первого социально-демографического типа ( более развитая часть рабочих и служащих - см. п. 5.2.I.).

Второй тип поведения, свойством которого является высокий удельный вес труда в домашнем хозяйстве, отличает сравнительно с остальными категориями обследованных более тесная связь с блоком таких характеристик, как проживание в малом городе, малоквалифицированный и неквалифицированный труд, низкая, менее I00 руб. заработная алата (левая часть рис.2).

Следовательно, данный тип поведения чаще всего сочетается со второй составляющей второго социально-демографичесного типа (менее развитая часть рабочих - см. п. 5.2.2.).

Четвертый тип поведения, с его существенно учебным характером, чаще компануется с молодежными группами, представители которых совмещают труд с учебой, имеют среднее образование, являются членами ВЛКСМ, т.е. с первой составляющей третьего социально-демографического типа (рабочая молодежь, п. 5.2.3.).

В этом смысле можно утверждать, что данные наборы характеристик представляют объективную социальную почву, условия, в которых возникают соответствующие типы времяпрепровождения.

Знаменательно, что связь поведенческих и социально-демографических групп, выявленная нами, кажется чрезвычайно логичной и позволяет лучше понять характер самих типов поведения, полученных при независимом анализе данных о времяпрепровождении.

Интересно также, что имеются поведенческие группы, не дающие или почти не дающие социально-демографических аналогий. Третий, "гармонический" тип поведения связан только с определенным возрастом ( 26 - 35 лет), т.е. более характерен для поколения рабочих, формирование которого пришлось на 50-ые годы. Пятый тип поведения практически не связывается ни с каким выявленным набором социально-демографических характеристик, что объясняется его особым пограничным характером. Очевидно формирование данных типов поведения зависит скорее от культурно-психологических факторов, чем от тех жизненных обстоятельств, которые отражаются в учтенных нами социально-демографических показателях.

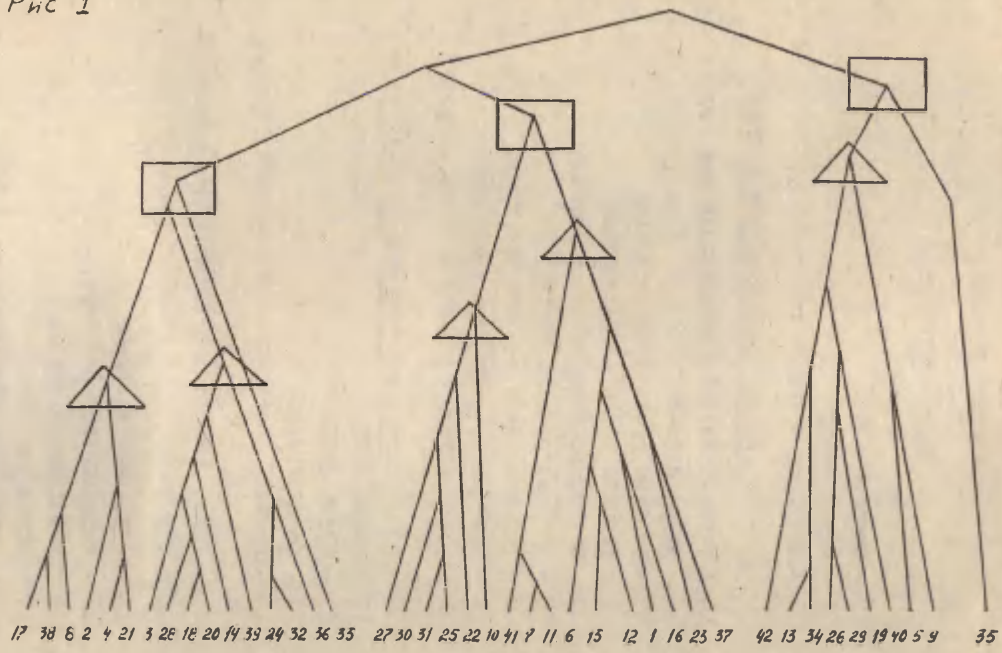
8. Наличие связи социально-демографических характеристик с основными типами поведения не дает, однако, ответа на вопрос о мере этой связи. Между тем, вопрос этот весьма важен, тем более, что обнаруживается не только связь,но и известная независимость некоторых сторон поведения относительно социально-демографических показателей. Данная

проблема требует специального рассмотрения и поэтому здесь мы ограничимся постановкой вопроса.

Мы считаем особенно важным подчеркнуть, что содержательная в социологическом смысле мера связи (зависимости) между социально-демографическими характеристиками и поведением может определяться через ошибку предсказания, прогноза поведения, полученного на основе объективных социально-демографических данных. Другими словами, мерой зависимости может служить степень совпадения прямого ( на основе данных и затратах времени) и косвенного ( на основе социально-демографических данных) распределения людей по типам поведения (времяпрепровождения). (Подобные методы измерения связи были изложены, в частности, в работе L. Goodman & W. Kruskall. Measures of association for classification. "Journ. of Amer. Stat.Ass.",1954, vol.49, p. 734 - 764.

Таким образом, если выделение относительно однородных типов и групп внутри исследуемой совокупности, равно как и принципиальный ответ на вопрос о наличии связи между рядами различных типов ( социальными типами и типами поведения), представляет собой задачу обратную распознаванию образов, то установленные меры связи между группами разного рода требует решения задачи распознавания образов в ее прямой, классической постановке.

Рис 1


17 38 8 2 4 21 3 28 18 20 14 39 24 32 36 35
27 30 31 25 22 10 41 7 11 6 15 12 1 16 23 37
42 13 34 26 29 19 40 5 9 35

## Пояснения к рис. I

### Социально-демографические характеристики, учтенные в анализе

#### Место жительства

I - большой город
2 - малый город
3 - пригород большого города

#### Квалификация и характер труда

4 - рабочие низкой квалификации или рабочие без квалификации
5 - рабочие средней квалификации
6 - рабочие высокой квалификации
7 - инженерно-технические работники
8 - рабочие ручного труда
9 - рабочие механизированного труда
I0 - низовые руководители (бригадиры)
II - руководители среднего и высшего уровня

#### Партийность и общественная работа

I2 - члены КПСС
I3 - члены ВЛКСМ
I4 - беспартийные
I5 - наличие постоянной общественной работы

#### Материальное положение и жилищные условия

I6 - отдельная квартира
I7 - коммунальная квартира
I8 - собственный дом
I9 - общежитие
20 - отсутствие коммунальных удобств
2I - заработок менее I00 руб. в месяц

22 - заработок I00 - I50 руб. в месяц
23 - заработок свыше I50 руб. в месяц
24 - месячный доход в семье менее 50 руб. на душу
25 - месячный доход в семье 50 - 75 руб. на душу
26 - месячный доход в семье свыше 75 руб. на душу
27 - наличие телевизора
28 - наличие подсобного хозяйства

Семейное положение

29 - несемейная молодежь
30 - семейные и пожилый люди
3I - наличие I несовершеннолетнего ребенка в семье
32 - наличие 2 несовершеннолетних детей в семье
33 - наличие 3 и более несовершеннолетних детей в семье

Возраст

34 - люди моложе 26 лет
35 - люди 26 - 35 лет
36 - люди 36 - 45 лет
37 - люди старше 45 лет

Образование и культура

38 - образование 4 класса и менее
39 - образование 5 - 7 классов
40 - образование 7 - I0 классов
4I - высшее, незаконченное высшее и среднее специальное образование
42 - совмещение труда с учебой.

## Пояснения к рис. 2

1 - 42 - социально-демографические характеристики (см. пояснения к рис. I).

I - принадлежность к первому (семейно-домашнему) типу времяпрепровождения

II - принадлежность ко второму (домашне-хозяйственному) типу времяпрепровождения

III - принадлежность к третьему (гармоническому) типу времяпрепровождения

IУ - принадлежность к четвертому (учебному) типу времяпрепровождения

У - принадлежность к пятому (остаточному) типу времяпрепровождения.

## SOTSIAAL-DEMOGRAAFILISTE TÜÜPIDE ERALDAMINE KLASTER-ANALÜÜSI MEETODITEGA JA NENDE SEOSE MÄÄRAMINE KÄITUMISTÜÜPIDE SUHTES

L. Gordon, E. Klopov, A. Terjohhin, M.Sivertsev
(Moskva)

Resümee

Töös lahendatakse klaster-analüüsi meetodiga sotsiaal-demograafiliste omaduste kogumi otsimise ülesanne, mis küllalt sageli esineb tegelikkuses. Esitatakse andmed 200 töölise sotsiaal-demograafiliste näitajate kohta, kes töötavad suurtes tööstusettevõtetes.

## THE DISTINGUISHING OF SOCIO-DEMOGRAPHIC TYPES BY CLUSTER-ANALYSIS METHODS AND THE DETERMINING OF THEIR RELATION TO BEHAVIORAL TYPES

L. Gordon, E. Klopov, A. Teryokhin, M.Sivertsev
(Moscow)

Summary

The problem of finding a set of socio-demographic characeristics is solved by means of cluster-analysis method that often occurs in practise. The data about socio-demographic characteristics of 200 workers in large plants are presented.

# К ВОПРОСУ О ПОСТАНОВКЕ ЗАДАЧИ ТИПОЛОГИИ

Н.В. Мартынова
(Свердловск)

Постановка вопроса о возможности использования методов распознавания образов при обработке информации конкретных социологических исследований является, на наш взгляд, вполне правомерной и своевременной.

В настоящее время уже имеется несколько довольно удачных попыток применения различных алгоритмов распознавания при систематизации социологической информации, но основное внимание в большинстве работ уделяется не методологической и методической стороне данного вопроса, а содержательной социологической интерпретации выделенных групп, классов, типов.

В связи с этим нам хотелось бы остановиться на некоторых методологических вопросах применения методов распознавания.

Относительно возможности использования методов распознавания при анализе социологической информации необходимо отметить, что здесь часто встречаются содержательные ситуации, аналогичные рассматриваемой в теории распознавания. Этот момент становится тем более понятным, если учесть, что теория распознавания образов является частью более общей теории принятия решений.

Однако на пути использования имеющихся методов распознавания, перенесения их из "точных" наук в общественные, в частности в социологию, имеется ряд затруднений. Остановимся на некоторых из них.

Первое затруднение состоит в самой постановке задачи типологии.

Мы не согласны с мнением некоторых исследователей, рассматривающих эти задачи как взаимно обратные:[I]

---

[I] Л.А. Гордон, В.Я. Волк, С.Е. Генкин, Э.В. Клопов, С.Н. Соколов. Типология сложных социальных явлений. "Вопросы философии", 1968, № 7.

На наш взгляд, между ними нет полного совпадения, но задача типологии в зависимости от цели исследования, от некоторых конкретных его особенностей может частично или полностью совпадать с задачей распознавания, именно с тем случаем, когда на первом этапе имеет место обучение "без учителя".

В связи с этим имеющиеся методы распознавания с учителем лишь частично могут быть использованы при решении задач типологии и особенно остро ставится вопрос о разработке методов распознавания "без учителя".

Вторая трудность в использовании известных методов распознавания при решении задач типологии состоит в низком уровне точности измерения признаков, привлекаемых при описании типов.

Остановимся несколько подробнее на постановке задачи типологии.

Мы рассматриваем типологию как более глубокий и полный метод систематизации по сравнению с группировкой и классификацией, как метод, тесно связанный с теорией изучаемого явления. При типологии социалистических объектов довольно часто, на наш взгляд, имеет место такая ситуация, когда исследователь, зная общее направление систематизации информации, определив основной интегральный признак типологии, интервал его изменения, не может указать перечень типов, правила отнесения объектов к тому или иному типу на основании теории явления и в решении этого вопроса большие надежды возлагает на ту информацию о типологии, которая содержится в эмпирическом материале.

При непосредственном наблюдении явлений, при анализе первичной информации перед исследователем последовательно проходят различные образы или описания этих образов. Если бы признаки, включенные в описание изучаемых объектов, были распределены случайно, то было бы невозможно выделить отдельные объекты и явления. Изучаемые информационные потоки (описания явлений) имеют определенную структуру в распределении признаков, характеризующуюся определенными

объективными закономерностями, внутренне присущими данной совокупности. К числу последних относится, например, наличие корреляционных связей между признаками, предопределяющих выделение определенных групп признаков или комплектов значений по этим признакам. Такие закономерности являются важной предпосылкой выделения типов.

При решении задач типологии, несмотря на их большое конкретное разнообразие, на наш взгляд, можно выделить следующие две постановки вопроса.

Во-первых, одной из задач систематизации информации является выделение типичных групп явлений, выяснение разнообразия типов, определение списка типов.

Во-вторых, весьма часто исследователь помимо определения разнообразия типов ставит задачу частичного или полного упорядочения их между собой. В этом случае правомерно отдельные типы представлять как стадии развития в определенном смысле исследуемого сложного явления.

Мы считаем, что главной задачей процесса типологии изучаемых объектов является не только констатация факта существования большего или меньшего разнообразия типов, а представление их как ступеней развития сложного явления и этим самым упорядочение их между собой, указание путей перехода от низких типов к высшим и выработка практических рекомендаций, связанных с этим переходом.

Относительно субординации двух выше сформулированных задач типологии в познавательном плане необходимо заметить, что они могут быть рассмотрены как различные уровни познания объектов или явлений, а именно, вторая задача дает больше информации исследователю по сравнению с первой. При этом выбор того или иного уровня изучения зависит не только от желания исследователя, но и от некоторых объективных условий исследования, в частности от содержания конкретной поставленной задачи.

Далее нам хотелось бы обратить внимание на некоторые стороны процесса типологии, имеющие методологическое значение:

А) Формулировка цели типологии, т.е. прежде всего должен быть решен вопрос о том, типологию чего мы проводим: или типологию общественной активности трудящихся, или типологию художественных интересов, или типологию телезрителей и т.д.

Выражение признака – цели через совокупность эмпирических референтов.

Б) Выбор признакового пространства, в котором будет проводиться типология. Дополнительно к признаку-цели необходимо определить признаки, в той или иной степени детерминирующие изучаемый объект, процесс.

В) Определение значимости отдельных признаков в процессе формирования типов.

Относительно этих трех моментов необходимо заметить, что первый из них в какой-то мере определяет второй и третий.

Выбор рациональной размерности признакового пространства пока является "узким местом" в методике конкретных исследований. Повышение размерности пространства, связанное с уточнением описания изучаемых объектов, выявлением их сущности, сопровождается быстрым ростом объема работы при анализе информации.

Исследование проблемы типологии показывает, что одной из важных и трудных ее ступеней является выделение существенной информации в описании изучаемых объектов и явлений. На языке многомерного представления объектов исследования эта задача может быть сформулирована как задача выделения комплекса существенных признаков. Мы особо подчеркиваем, что решение этого вопроса неотделимо от постановки самой цели исследования. Если не определена ц е л ь типологии, то отсутствуют и критерии для определения существенности признаков, их значимости в процессе познания данного явления.

Исследователь, определяя цель исследования как некоторый сложный признак, обычно представляет его через систему более простых признаков, которые мы предлагаем называть признаками-критериями типологии. Представляет интерес изучение всех возможных сочетаний знаний этих признаков-критериев. При типологии в первую очередь должны быть учтены те сочетания, кото-

рые представляют теоретический интерес. Кроме того при анализе необходимо учесть сочетания, отличающиеся своей "необычностью", например, сочетания, отражающие какие-то новые, зарождающиеся или наоборот старые, отживающие тенденции в развитии явления, несмотря на их относительную малочисленность.

Вопрос о принципах выделения "интересных" мест в процессе типологии явлений заслуживает серьезного внимания, и, по-видимому, в определенной степени обеспечивает успешность применения этого метода систематизации информации.

Мы считаем, что к таким принципам (правилам) можно отнести:

1) анализ всех возможных сочетаний значений признаков-критериев типологии,

2) оценку теоретической значимости отдельных сочетаний,

3) критерий массовости, репрезентативности сочетаний.

Эти критерии далеко неравноценны и при анализе информации эту неравномерность необходимо учитывать. В связи с этим нам кажется неправомерным выдвижение на первое место критерия м а с с о в о с т и и использование его как основного, а тем более единственного. Хотя применение аппарата таксономии при типологии, основанного только на этом принципе, в ряде случаев дает не плохие результаты, с методологической точки зрения, данный критерий не может стоять на первом месте, его основная функция - корректирующая. Начинать типологию, очевидно, нужно с анализа теоретически значимых сочетаний знаний признаков-критериев, более точно определяющих сущность изучаемого объекта или процесса .

Интересным и заслуживающим внимания моментом является в о п р о с о т и п о л о г и ч е с к и х с в о й с т в а х.

Процессу типологии предшествует изучение отдельных свойств, в той или иной мере привлекаемых в процессе изучения объектов и явлений. Б.М. Теплов, занимавшийся проблемами типологии высшей нервной деятельности, писал: "Душа учения о типах высшей нервной деятельности - проблема типологических свойств нервной системы. Успешность разработки проб-

лемы типов зависит в первую очередь от достигнутой глубины понимания природы типологических свойств"[2]. Эта мысль, на наш взгляд, полностью справедлива и для социологии.

Какие признаки являются типологическими? В имеющихся у нас работах по типологии этой проблеме не уделяется должного внимания. Обычно все признаки, хотя бы в какой-то степени привлекаемые в процессе типологии, называют типологическими. Следствием этого является тот факт, что такие признаки как пол и возраст оказываются типологическими при изучении почти каждого социологического явления. Мы считаем это неверным, отнюдь не вытекающим из самой сути метода типологии.

На наш взгляд, признаки, привлекаемые при типологии социологических объектов и явлений, целесообразно разделить на две группы:

I. группа признаков-критериев, II группа признаков-детерминант. Это связано со следующими соображениями.

Как правило, метод типологии применяется при изучении сложных социальных явлений и сущность его состоит в выделении типов на основе анализа некоторых интегральных характеристик этих явлений, определяемых целью исследования. Эти интегральные характеристики часто носят скрытый, латентный характер и изучение их, в частности, определение интервала изменения, представляет трудную самостоятельную задачу.

При сложившемся многокачественном подходе при изучении сложных объектов и процессов изучение самого объекта заменяется изучением системы его свойств. Чем глубже разработана теория изучаемого явления, тем точнее может быть определена система свойств, признаков, отражающих сущность исследуемого феномена.

В самом общем плане цель любой типологии можно рассматривать как некоторый сложный интегральный признак, сущность которого проявляется через систему его признаков-проекций.

---

2) Б.М. Теплов. Проблемы индивидуальных различий. М., изд-во АПН, 1961, стр. 479.

Именно эти более простые признаки, компоненты интегрального признака - цели, мы относим к I-ой группе, группе признаков-критериев.

Так, при изучении общественной активности трудящихся в эту группу нами были включены такие признаки: общее отношение к общественной деятельности, отношение к конкретному общественному поручению, временные затраты, инициатива, число выполняемых поручений. Эти признаки-критерии позволяют составить представление о степени общественной активности трудящихся в целом и отдельных групп.

Во вторую группу были включены социально-демографические факторы, материальные условия жизни, признаки, связанные с производственной деятельностью людей и т.д., т.е. признаки, связанные с причинной обусловленностью наличия различных степеней общественной активности и с условиями ее проявления. Есть все основания ожидать, что при соответствующих изменениях этих условий будет изменяться и тип активности личности. Однако , эти признаки не могут быть положены в основу выделения типов общественной активности, так как они не отражают сущности этого феномена. Основная функция этой группы признаков-детерминант - объяснить наличие имеющегося разнообразия типов, охарактеризовать предпосылки появления отдельных типов.

Как следует из вышесказанного, выделяемые нами группы признаков играют различную содержательную роль в процессе типологии. I группа признаков-критериев определяет содержание изучаемого феномена, шкалу этого сложного признака, его отдельные качественно-определенные состояния. Именно система этих градаций сложного признака предопределяет разнообразие типов, т.е. по существу определяет результат процесса типологии. Поэтому признаки этой группы мы предлагаем считать типологическими. С их изучения и начинается типология.

В соответствии с выделением двух различных групп признаков мы считаем, что в процессе типологии целесообразно также выделить два этапа.

I этап - проведение типологии изучаемых объектов и явлений на основании признаков-критериев.

а) Построение производной шкалы интегрального признака типологии на основании признаков-критериев; определение интервала изменения этого признака; квантификация его, в частности, выделение интервалов значений, соответствующих различным типам.

При построении производной шкалы интегрального признака в зависимости от характера отношений между признаками-критериями могут использоваться различные методы, в частности, методы таксономии. Мы в своем исследовании использовали описанную в предыдущем параграфе процедуру сведения зависимых признаков.

б) Разбиение множества наблюдений на подмножества, соответствующие различным типам, т.е. определение поэлементного состава типов, производящееся в строгом соответствии со значением интегрального признака, его мерой.

Если шкала построена, выбирается определенный интервал значений сложного признака или соответствующие ему наборы значений комплекса $n$ признаков-критериев и отбираются все наблюдения, попавшие в этот интервал. Затем выбирается следующий интервал значений и повторяется процедура отбора.

II этап - описание выделенных подмножеств наблюдений с привлечением $(N - n)$ признаков[3] II группы. II этап типологии не связан, на наш взгляд, с какими-либо серьезными изменениями в структуре выделенных типов, в их упорядочении между собой. Основное его назначение состоит в объяснении наличия системы типов, в определенном обосновании ее.

Типология, предложенная на I этапе, должна быть подтверждена, объяснена и, возможно, скорректирована при

---

[3] $N$ - общее число признаков I и II групп.

использовании более широкого списка признаков, дополненного за счет признаков II группы.

Предварительное выделение типов на основании теоретических предпосылок и частичного использования эмпирического материала, а затем подтверждение и корректировка типологии за счет рассмотрения дополнительных признаков должно способствовать более точному решению проблемы типологии.

В общем виде I этап расчленения множества точек $n$-мерного пространства на подмножества на основании признаков-критериев можно представить так:

а) представление целевой функции (интегрального признака типологии) через систему признаков-критериев - построение модели; определение области значений целевой функции; выделение качественно-определенных состояний (градаций) по интегральному признаку; формирование описаний типов, т.е. определение для каждого типа интервала значений по одномерному признаку-цели с одновременным описанием типа через систему $n$ признаков-критериев.

б) задается определенное значение или интервал значений целевой функции (этим определяется система интервалов значений по признакам-критериям), затем из множества наблюдений выбираются те, в которых целевая функция принимает заданное значение или значение из указанного интервала. Таким образом, все множество наблюдений делится на подмножества (типы), в которых целевая функция ( интегральный признак типологии) принимает вполне определенные значения. Типы, выделенные таким способом, оказываются упорядоченными по признаку-цели типологии и частично упорядоченными по всем или некоторым признакам-критериям.

С более формальной точки зрения выделение в процедуре типологии двух этапов можно интерпретировать следующим образом.

Если каждое наблюдение представить как точку $k$-мерного признакового пространства и задачу типологии формулировать как разбиение множества точек, заполняющих

пространство, на подмножества, соответствующие отдельным типам в смысле признака-цели, то I этап типологии равносилен поиску этих подмножеств в наиболее информативном подпространстве (размерности $r < h$) этого пространства, а именно, в подпространстве признаков-критериев.

Успешность решения задач распознавания образов, методами которых могут решаться задачи типологии, определяется используемой в данном пространстве метрикой. Определение же метрики равносильно нахождению системы весов признаков, их значимости в раскрытии содержания интегрального признака, положенного в основу типологии. Чем точнее будет определена система весов, тем правильнее будет типология. Этот факт необходимо иметь в виду и при типологии социологических объектов и явлений.

Ввиду того, что социолог чаще всего не имеет ни системы весов используемых признаков, ни методов их определения, т.е. решение этой проблемы в полном объеме невозможно, нам кажется, что в практике конкретных исследований полезно использовать любые частичные решения этой проблемы. Мы считаем, что выделение в процессе типологии двух этапов и первоначальный поиск типов в наиболее информативном подпространстве признакового пространства (I этап типологии) является частичным решением этой проблемы. При этом исследователь фактически $h$ признаков, привлекаемых при типологии, разбивает на две группы по их информативности:

$r$ признаков-критериев - более информативная группа;

$(h-r)$ признаков-детерминант - менее информативная группа.

Необходимо заметить, что такое деление можно продолжать дальше. Ввиду того, что признаки-детерминанты также неравноценны между собой в определении типов, мы предлагаем выделить в особую группу на этом этапе признаки-причины изучаемого процесса. Именно с наличием этой группы признаков мы связываем объяснительную функцию типологии.

Таким образом, общий список признаков, привлекаемых при типологии тех или иных объектов, может быть из чисто содержательных соображений поделен на следующие группы:

1) признаки-критерии,
2) признаки-причины
3) остальные признаки-детерминанты } признаки-детерминанты,
4) группа признаков, характеризующих фон исследования (начальные условия-система принятых исходных допущений).

В руководствах, монографиях, статьях, посвященных проблеме распознавания образов, в частности, в статистической теории отмечается, что одним из труднейших моментов этого процесса является составление характеристик "образов", "типов". Этот этап в теории распознавания образов принято называть обучением. Мы считаем, что процесс поиска типов, в некотором смысле, совпадает с первым этапом распознавания - обучением. Причем, это обучение без учителя, т.е. список типов, классов априори здесь неизвестен.

Как отмечает Н.Нильсон[4], задача распознавания образов с учителем самая разработанная в настоящее время, но не единственная и не самая интересная среди задач, возникающих в этой области. Внимание исследователей все больше привлекают следующие две задачи: обучение распознаванию образов без учителя и составление описания сложного конкретного изображения, явления. На наш взгляд, большое число проблем социологических исследований может быть сведено именно к этим двум типам задач, в частности - проблема типологии.

Типология социальных явлений, определение списка типов и их характеристик, по-видимому, не всегда является конечной целью исследования. Если поставить вопрос несколько шире: "Для чего мы получаем описания типов?" - то здесь и появляется второй этап задачи распознавания образов. Одним из практических выходов типологии являет-

[4] Н.Нильсон. Обучающиеся машины. М., "Мир", 1967, стр. 8.

ся именно распознавание социологических объектов. когда типология по какому-либо сложному признаку, полученная в одном исследовании, при соответствующих условиях используется в других.

Например, исследователя, не занимающегося специально проблемой общественно-политической активности, в связи с изучением других проблем может интересовать уровень общественной активности исследуемого контингента. Для этого ему достаточно познакомиться с конкретными условиями проявления этой активности. Если исследователь знаком с типологией общественной активности т.е. знает социальную характеристику каждого типа (описание уровней активности и социальные характеристики типов), то на основании социального портрета группы и условий проявления активности он может с определенной достоверностью воспроизвести и уровень активности.

Таким образом, практически решается задача отнесения вновь исследуемого индивида или группы индивидов к одному из уже ранее определенных типов общественно-политической активности на основании некоторых характеристик II группы. Это и есть II этап задачи распознавания образов, "собственно распознавание". Но эта часть проблемы обычно не ставится на первом этапе исследования типов и вообще обычно остается в тени, очевидно, в связи с большими трудностями этого I этапа.

По-видимому, нельзя слишком категорично противопоставлять задачи типологии и распознавания образов. Мы считаем, что : I) если типология является конечной целью исследования и дальнейшее ее использование не планируется или невозможно в принципе, то типология полностью совпадает с первым этапом задачи распознавания - обучением без учителя; 2) существуют такие постановки задач, когда типология является лишь первым этапом, а общая задача ставится шире - как задача распознавания образов.

# TÜPOLOOGIAÜLESANDE PÜSTITAMISEST

N.V. Martõnova
(Sverdlovsk )

Resümee

Artikkel on pühendatud rühmitamise, tüpoloogia ülesannete mõningatele metodoloogilistele küsimustele.

Autor märgib tüpoloogia ülesannete osalist või täielikku kattumist õpetajata kujundite eristamise ülesannetega, mis tingib just selle ülesannete tüübi metodoloogia läbitöötamist. Märgitakse ka tunnuste mõõtmisel esinevate madalast täpsusest tingitud raskusi.

Tüpoloogiat loeb autor sügavamaks ja täielikumaks süstematiseerimise meetodiks rühmitamise ja klassifitseerimisega võrreldes, viidates selle tihedamale seosele uuritava nähtuse sisuga.

Tüpoloogia ülesannete lahendamisel eristab autor järgmisi küsimusi.

1) Tüüpide loetelu koostamine.

2) Tüüpide (osaline või täielik) järjestamine; eriti vajalik on see tüüpide käsitlemisel arengustaadiumitena.

Selleks on tarvis:

A) formuleerida tüpoloogia eesmärk;

B) valida tunnuste ruum;

C) määrata tunnuste olulisus.

Sotsioloogiliste objektide ja nähtuste kirjeldamisel kasutatavad tunnused jaotatakse kahte rühma - kriteerium-tunnused ja determinant-tunnused.

Tuuakse konkreetne näide mõlemat tüüpi tunnuste kohta töötajate ühiskondliku aktiivsuse uurimisel.

Tüpologiseerimise protsessis eraldatakse 2 etappi:

I - tüpologiseerimine kriteerium-tunnuste alusel;

II - saadud alamhulkade kirjeldamine determinant-tunnuste abil.

Käsitletakse ka mõningaid küsimusi seoses ruumi meetrika, tunnuste kaalude, informatiivse tunnuste alamhulga leidmisega.

## ON RAISING THE TYPOLOGY PROBLEMS

N. V. Martynova
(Sverdlovsk)

Summary

Some methodological issues connected with grouping typology problems are presented in this paper.

The author observes the partial or total coincidence of typology problems with pattern recognition problems without the 'teacher' that being the cause for elaboration of corresponding methodology. Some difficulties arise due to insufficient precision in measuring the variables.

The author considers typology to be a more profound and perfect method than grouping or classification referring to the fact that typology is more closely related to the essence of phenomenon than other methods.

In solving the typology problems the author distinguishes the following stages:

1) The making up of list of types

2) The ranking (partial or total) of types.

It has a special necessity in case when types are treated as developmental stages.

Together with this we must

A. formulate the aim of typology

B. choose variables' room

C. determine the relevance of variables.

The variables describing sociological objects and phenomena are divided into two groups: the criterion-variables and the determinant-variables.

An example is given about the occurrence of both types of variables in investigating the workers' social activity.

In the process of typologizing two stages are brought out:

I on the basis of criterion-variables

II the given subsets are described by means of determinant-variables .

Some issues connected with room metrics,the weighing of variables, the finding of informative subset of variables is also treated.

# ОПЫТ ПРИМЕНЕНИЯ МАТЕМАТИЧЕСКОЙ ПРОГРАММЫ ДИФФЕРЕНЦИАЛЬНОЙ ДИАГНОСТИКИ ДЛЯ ПРЕДСКАЗАНИЯ УСПЕВАЕМОСТИ СТУДЕНТОВ ВУЗА

А.Я. Левин

(г.Горький)

Редкая встреча социологов с представителями точных или прикладных наук обходится без того, чтобы кто-либо из последних, выслушав социологические доклады и дискуссии, не произнес бы недоуменно: "Ну и что?"

Многое можно сказать для обоснования права социологов заниматься вопросами, не имеющими непосредственного выхода в практику. Однако несомненно - престиж конкретной социологии зависит прежде всего от ее способности прогнозировать поведение людей и давать на этой основе рекомендации для социального управления. Как раз при решении такого рода прикладных задач, где необходимо устанавливать многофакторные зависимости на ограниченном статистическом материале, метод распознавания образов является наиболее мощным, а во многих случаях и единственным средством анализа социологической информации.

Нам представляется, что задача прогнозирования поведения различных индивидов в определенной ситуации, с формальной точки зрения, аналогична задачам дифференциальной медицинской диагностики, с которыми, как Ю.И. Неймарк показал в своем докладе, математики научились успешно справляться. Для проверки этого предположения была сделана попытка предсказать результаты экзаменов студентов ряда факультетов Горьковского университета по предметам, которые дают систематически наибольшее число неудовлетворительных оценок.

В качестве исходного материала использовались анкеты, собранные в ходе исследования причин неуспеваемости студентов в 1966 - 1968 гг. [1] Ответы на вопросы анкеты содержали сведения об оценках в школьных аттестатах, участии

в математических, физических и химических олимпиадах для школьников, о результатах предыдущих экзаменационных сессий, бытовых условиях студентов. Ряд вопросов анкеты позволял выявить мнения студентов о ходе экзамена, систематичности собственной работы во время семестра, качестве своей подготовки и пр. После соответствующей переработки данные, содержавшиеся в анкетах, были закодированы в 33 признаках, в каждом признаке было от 2 до 5 градаций. В первичном анализе анкетного материала и его приспособлении к требованиям математической программы принимали участие Р.И. Никифоров и А.А. Терентьев.

По программе, разработанной и неоднократно успешно использованной для решения задач медицинской диагностики Ю.И. Неймарком и его сотрудниками [2], З.С. Баталова и М.Г. Аранович провели анализ собранной информации. К сожалению, материал оказался недостаточен для получения статистически достоверных результатов ( из всех собранных анкет для машинного анализа оказались пригодны только 137). Однако принципиальная ценность метода распознавания образов для решения социологических задач, в частности, возможность использования для этой цели программ дифференциальной медицинской диагностики были, как нам представляется, подтверждены.

Если считать удовлетворительным прогноз, расхождение которого с действительной экзаменационной оценкой не превышает одного балла, то его точность составила 0,89. Хороший прогноз (полностью совпадающий с экзаменационной оценкой) составляет 0,49. Наиболее точен прогноз отличной оценки. Он равен соответственно - 0,95 и 0,64.

Возможность предсказания оценок наглядно показывает , что по успеваемости студенты делятся на сравнительно четко разграниченные группы, в каждой из которых диапазон оценок почти не выходит за пределы двух смежных. Особенно резко выделяется группа отличников. Существование этих групп ощущается и молчаливо признается преподавателями вузов. Однако имплицитативного признания недостаточно, По-

скольку такие группы являются реальностью, то их существование необходимо открыто и постоянно учитывать в методике преподавания, в учебных планах, в определении специализации, в выдаваемых дипломах и, что особенно важно, в распределении выпускников на работу.

Использованная программа позволяет вычислять веса различных признаков, отражающие их сравнительное значение в отнесении объектов к той или иной группе. В данном случае веса всех признаков не вычислялись. Было определено значение только некоторых из них. Выяснилось, что, как и следовало ожидать, наибольший вес в прогнозе принадлежит оценкам предшествовавших сессий. Определенное значение имеют также оценки школьного аттестата. Последнее обстоятельство свидетельствует в пользу уже выдвигавшегося предложения об освобождении абитуриентов, имеющих в школьных аттестатах отличные оценки по соответствующим дисциплинам, от вступительных экзаменов по непрофилирующим предметам. Выяснилось также, что группа признаков, основанных на мнениях студентов о качестве своей подготовки к экзамену, систематичности собственной работы во время семестра и пр., то есть вся информация, которую можно отнести к разряду "субъективной", не имеет существенного значения для прогноза успеваемости.

Выше уже говорилось, что работа предпринималась, главным образом, с методической целью и показала пригодность программы дифференциальной диагностики для решения социологических задач определенного класса. Кроме этого основного вывода, можно сделать еще ряд замечаний методического характера.

Один из выступавших здесь докладчиков заметил, что проблемы классификации и определения весов - в числе самых сложных для социолога. Как нам представляется, преимущества метода распознавания образов как раз и состоят в том, что на его основе возможно создание программы, которая дает точные веса практически любого числа признаков и, следовательно, позволяет с наибольшей определенностью установить границы групп.

Понятно, что этот результат может быть получен только при правильной постановке социологической задачи. Прибегая к помощи математики, социолог должен ясно понимать, к чему это его обязывает. В частности, при использовании распознавания образов для установления весов признаков следует иметь в виду, что этот метод может быть полезен только в тех случаях, когда исследователь объединяет индивидов в группу по общности их поведения или общности результатов их деятельности в определенных, строго фиксированных ситуациях. Поэтому выбор такой социально значимой ситуации и четкое определение тех градаций поведения, которые будут использоваться в качестве критериев классификации, - непременные условия применения соответствующих программ.

Следующее замечание относится к сравнительному значению признаков, основанных на "объективной" и "субъективной" информации. Нам представляется, что незначительный вес последних в прогнозе успеваемости не случаен. Это, видимо, еще раз свидетельствует, что вербальное и реальное поведение образуют две слабо связанные между собой системы. Отсюда следует, что прогноз реального поведения лишь ограниченно может полагаться на оценочные высказывания изучаемых индивидов и социолог ( во всяком случае при обычной для современных исследований технике сбора информации)должен опираться, главным образом, на данные о реальном поведении этих индивидов в прошлом.

В заключение - о влиянии, которое использованием математических программ, основанных на методе распознавания образов, может оказать на методологию самого социологического исследования. До сих пор наиболее существенные результаты были получены в работах с тщательно сформулированной гипотезой, так что само исследование должно было либо подтвердить эту гипотезу, либо опровергнуть ее. Чаще всего таким путем устанавливается значение одного какого-либо фактора. Преимущества здесь вытекают из возможности сконцентрировать усилия на решении определенной задачи и получить ответ именно на тот вопрос, который интересует иссле-

дователя. Вместе с тем не следует закрывать глаза на некоторую ограниченность возможностей этой процедуры.

Часто для социолога, особенно в прикладных задачах, важно не столько установить сам факт связи, сколько сравнить между собой влияние множества различных факторов. В таких работах только сопоставление весов позволяет построить многофакторное объяснение, учитывающее как типичные социальные признаки, так и специфические личностные синдромы. Подчас поэтому целесообразно заложить в программу все те сколько-нибудь значимые признаки, которые могут быть учтены и измерены. Результаты здесь могут оказаться неожиданными для самого исследователя и, естественно, не могли бы быть предусмотрены в гипотезах, которые он способен построить. Именно для такой процедуры метод распознавания образов является адекватным средством анализа.

## ЛИТЕРАТУРА


I. Социологические исследования учебно-воспитательной работы в высшей школе. Ученые записки Горьковского государственного университета, вып.9I, Горький, I969, стр. I68 - I70.

2. Неймарк Ю.И., З.С. Баталова. Опыт использования быстродействующей вычислительной машины для медицинской диагностики, прогнозирования исхода оперативного вмешательства или заболевания и выбора оптимального метода лечения. В кн.: Прикладная математика и кибернетика. Материалы к всесоюзному симпозиуму по прикладной математике и кибернетике. Ученые записки, Горький, I967, стр. 293 - 369.

## KATSE ENNUSTADA ÜLIÕPILASTE ÕPPEEDUKUST DIFERENTSIAALDIAGNOSTIKA MATEMAATILISE PROGRAMMI ABIL

A.J. Levin (Gorki)

Resümee

Erinevate indiviidide teatud situatsioonis käitumise ennustamine on formaalselt analoogiline meditsiinilise diferentsiaaldiagnostikaga,mille matemaatilised alused on juba välja töötatud. Selle hüpoteesi kontrollimiseks prooviti ennustada rea Gorki RÜ teaduskondade üliõpilaste eksamihindeid (ainetes, kus esineb suhteliselt palju mitterahuldavaid).

Lähtematerjaliks olid 33-tunnuselised ankeedid,mis sisaldasid koolitunnistuste ja eelnevate sessioonide eksamihindeid; Neimarki jt. diagnostika programmi alusel analüüsiti 137 ankeeti.

Tegelikud eksamitulemused langesid prognoositavatega kokku 49%-l juhtudest,erines ülimalt ühe hinde võrra 89%-l juhtudest.

## AN ATTEMPT TO PREDICT STUDENTS' SCHOLASTIC EFFECIENCY ON THE BASIS OF A MATHEMATICAL PROGRAM IN DIFFERENTIAL DIAGNOSTICS

A.Y.Levin (Gorki)

Summary

The prediction of an individual's behaviour in a situation is formally analogous to differential diagnostics in medicine the mathematical fundamentals of which have been worked out. To test this hypothesis an attempt was made to predict the marks several students of the Gorki State University got when examined in subjects where unsatisfactory marks prevailed.

The analysis was based on the data got from questionnaires of 33 variables comprising the marks of school reports and previoris sessions. 137 questionnaires were analysed on the basis of Neimark diagnostics program.

The actual marks coincided with predicted ones in 49 per cent of cases; the difference in one mark being in 89 per cent of cases.

# АБСТРАКТНАЯ ЖИВОПИСЬ КАК ОСОБЫЙ - ВЫРОЖДЕННЫЙ ЯЗЫК. (СТАТИСТИЧЕСКИЕ МЕТОДЫ ИЗУЧЕНИЯ ВОСПРИЯТИЯ АБСТРАКТНОЙ ЖИВОПИСИ).

П.Ф. Андрукович, В.С. Грибков, В.П. Козырев,
В.В. Налимов, А.Т. Терехин.
(Москва)

## I. Логический анализ языка абстрактной живописи.

В литературе уже не раз высказывались суждения о том, что абстрактная живопись может рассматриваться как знаковая система, представляющая собой особый язык (см., например, [1] ). Цель этой работы - подвергнуть данное утверждение четкому логическому анализу, широко используя математические методы для изучения суждений экспертов, которым предлагалось по определенным правилам читать произведения абстрактной живописи.

Смысловое содержание понятия языка расширяется по мере того, как, с одной стороны, создаются новые искусственные языки, с другой стороны, включаются в рассмотрение все новые и новые, подчас совсем необычные информационные системы. Мы надеемся, что рассмотрение абстрактной живописи, явления интересного и до конца не понятого, также обогатит наше представление о языке.

Нам кажется, что сейчас нельзя дать достаточно хорошее определение понятия языка, но можно, по крайней мере, охарактеризовать его основные свойства. Первой такой характеристикой должна быть функциональная характеристика. Язык функционирует как средство общения [2]. Это есть некоторая система, служащая

для передачи какой-то информации [1]. Абстрактная живопись несомненно выполняет коммуникационную службу, передавая какую-то информацию от художника к зрителю. Во всяком случае, здесь явно имеется система, состоящая из передатчика - художника, приемника - зрителя и средства общения между ними. Но все же остается не ясным, можно ли эти средства общения классифицировать как языковую систему. Многочисленные горячие и не всегда корректно проводимые споры показывают, что такая постановка вопроса вполне правомерна.

Чтобы ответить на поставленный выше вопрос, попробуем, следуя Ю.А. Шрейдеру [3], рассмотреть структуру языка. Прежде всего, язык должен состоять из некоторой системы знаков и синтактики - правил комбинирования знаков. Правда, мы не умеем сейчас сколько-нибудь удовлетворительно определить понятие знака и знаковой системы, полагая, что на интуитивном уровне смысл этих понятий воспринимается достаточно хорошо. Совокупность всех знаков образует алфавит системы. Знаки, скомбинированные по определенным правилам, создают тексты. Изучение знаковых

---

[1] Термин "информация" мы также не можем определить, он достаточно хорошо воспринимается и не требует дополнительных пояснений. У читателя не должно вызывать удивления то обстоятельство, что мы пытаемся проводить логический анализ, не определив основные понятия. Известный математик Гильберт, даже при построении строго формализованных систем не сумел дать определения всем нужным ему понятиям; многие из них определились только через те аксиомы, которые с их помощью формулировались.

систем может проводиться в трех направлениях: I) синтаксическом, когда определяются правила упорядочения знаков в текстах, 2) семантическом, когда определяется информация, закодированная в знаковой системе и,таким образом,выявляется отношение текста к "передатчику", 3) прагматическом, когда исследуется содержание, извлеченное из текста, и таким образом определяется отношение текста к "приемнику". Одна из особенностей языка состоит в том, что один и тот же язык может быть представлен в различных знаковых системах, образующих некоторую иерархическую систему разных уровней. Например, для обыденного языка, скажем, для русского языка мы имеем систему уровней, состоящую из букв, морфем [2], словоформ [3], сегментов [4], фраз [5] и т.д. Ю.А.Шрейдер в упомянутой выше работе [3], считает, что это свойство языка может служить его определением. В его терминологии это звучит так: "языком мы будем называть категорию эквиморфных знаковых систем".

Попробуем теперь с рассмотренных выше позиций подвергнуть анализу абстрактную живопись. Прежде всего, нам нужно рассмотреть первичную систему знаков абстрактной живописи, находящуюся на нижнем уровне иерархии знаковой системы. В приложении I приведен список таких первичных знаков, образующих алфавит абстрактной живописи. Возможно, что этот список далеко не полон.

---

[2] Морфема - значимая часть слова: корень и аффикс (приставки, суффиксы и пр.).

[3] Словоформа - отрезок текста между пробелами.

[4] Сегмент - отрезок текста между знаками препинания.

[5] Фраза - отрезок текста между точками.

Важно подчеркнуть, что эта система знаков не поддается разложению на знаки более низкого уровня. Интересно также отметить, что система знаков оказывается достаточно богатой в смысле своего разнообразия. Может быть, самым важным здесь оказывается то обстоятельство, что эта система знаков оказывается в информационном смысле п у с т о й : им нельзя поставить в соответствие какую-либо информацию.

В связи с этим интересно рассмотреть вопрос о том, является ли обязательным свойством языка строгое соответствие между знаками и той информацией, которая кодируется этими знаками, и есть ли примеры существования вырожденных - совершенно пустых знаковых систем. Если мы возьмем буквенный алфавит обыденных (естественных) языков людей, то здесь легко прослеживается довольно строгое соответствие между буквами и фонемами [6], которые кодируются буквами. Если же мы поднимемся выше по иерархической лестнице и будем рассматривать словоформы, или попросту - слова обыденного языка, то здесь уже не будет простого, однозначного соответствия между словом и той информацией, которая кодируется словом. В англо-американской лингвистической литературе принято различать две системы кодирования информации словом / 4 /. Можно говорить о концепции с о о т н о ш е н и я (referent). Слово относится к определенному объекту или нескольким объектам. Это свойство слова определяется более или менее четко. Соотношение отнесения создает лишь бедный язык - люди идут дальше и приписывают словом особый смысл (meaning). Ут-

---

[6] Фонема - минимальная единица звукового строя обыденного языка; фонема служит для складывания и различения морфем.

верждается, что смысл слова черпается изнутри сознания человека. Слово есть некий "черпак", единый для всех, но у разных людей содержимое этого черпака оказывается различным [7]. В английской лингвистической философской школе обращается особое внимание на полиморфность (или полисемию) обыденного языка и утверждается, что как раз благодаря полиморфности естественный язык сильнее любого искусственного строго формализованного языка [5]. При этом полиморфность естественного языка отнюдь не задается такими простыми средствами, как синонимия или омонимия. Полиморфность обыденного языка возникает потому, что мы используем нечеткие и неотчетливые слова с неровными краями и неясными разграничительными линиями. Многообразие, допускаемое при кодировании и декодировании речевого поведения позволяет в вежливой форме, не раздражающей собеседника, нарушать узость строго дедуктивных форм мышления [8]. Этим, собственно, и опре-

---

[7] Иллюстрируем это утверждение примером, заимствованным также из [4]. Возьмем слово "минута" - оно имеет четкое соотношение - соответствует астрономической единице времени. Но вот представьте себе, что кто-нибудь собирается в театр, и его подруга говорит: "Подождите, минуточку". Эта совсем простая фраза имеет для разных людей различный смысл. Для кого-то это значит, что можно уже не торопиться, все равно опоздает на первое действие. Во всяком случае, в этом контексте слово "минутка" уже не имеет никакого отношения к астрономической единице времени.

[8] Речевое поведение должно иметь несколько антитез, иначе оно будет напоминать фальсифицированные судебные процессы с заранее предрешенным исходом.

деляется богатство эстественного языка. В [6] сделана попытка обосновать преимущества полиморфного языка, исходя из теоремы Гёделя о неполноте. Хорошо известная в математической логике теорема Гёделя (1931 г.) несет очень большую, еще не понятую до конца гносеологическую нагрузку. Один из общефилософских выводов, следующих из этой теоремы, может быть сформулирован примерно так: интеллектуальная деятельность человека богаче строго дедуктивных форм мышления [9] . Трудно сейчас с уверенностью сказать, как происходит творческое мышление, и хотя при коммуникации, на уровне языкового поведения, люди должны быть логичны, однако эта логичность не должна быть чрезмерно строгой - что достигается полиморфностью языка.

Итак, мы видим, что требование однозначности в кодировании и декодировании знаковых систем не является обязательным требованием к структуре языка. Более того, нарушение этой однозначности создает языки более сильные, чем строго формализованные искусственные языки. Правда, при этом возникает вопрос: сколь далеко может идти такое нарушение однозначности? В обыденном языке людей как-то спонтанно устанавливается некоторая разумная граница полиморфности, хотя иногда она, по-видимому, нарушается так, как это, скажем, имеет место сейчас в философии. Во всяком случае, слишком сильное нарушение правил кодирования в естественном языке приведет уже к ситуации психиатрической лечебницы. Первичная знаковая система абстрактной живописи является пустой, т.е. там нет никаких правил, связывающих эти знаки с какой-либо

---

[9] Глубоко осмысленное с общефилософских позиций и в тоже время вполне доступное изложение теоремы Гёделя можно найти в [7] .

информацией, которая может передаваться с помощью этих знаков. Это крайний, вырожденный случай языковой знаковой системы. Любопытно отметить, что в строго формализованных системах, скажем, в языках логики, также могут быть пустые знаковые системы, но при этом там строго формулируются правила оперирования со знаками.

Перейдем теперь к анализу синтактики. Выше мы отметили уже, что в строго формализованных языках, скажем, в языках математической логики, имеют место строго определенные правила оперирования со знаками. В обыденных языках эти правила задаются грамматикой и она оказывается совсем не строгой. Точнее, нужно было бы сказать так: построенные сейчас грамматики естественных языков не задают полностью языковые структуры. Эти грамматики оказались совсем неудовлетворительными с позиций машинного перевода [3] . Сейчас делаются попытки построения новых, более сильных грамматик для естественных языков, но не ясно, увенчается ли эта задача успехом, так как никто еще не доказал, что построение строго формализованной грамматики для естественных языков есть разрешимая задача. Во всяком случае ее разрешимость представляется весьма сомнительной, если принять бегло изложенную выше концепцию об оптимальности полиморфных языков, которая может быть обусловлена не только смысловой многозначностью слов, но и нечеткостью грамматик. Вернемся теперь к языку абстрактной живописи. В том же приложении I приведена некоторая грамматика абстрактной живописи - список простейших правил оперирования первичными знаками в картинах. Эта грамматика составлена феноменологически - путем анализа небольшой коллекции картин; список грамматических правил, видимо, далеко не полон.

Здесь важно опять-таки подчеркнуть, что приведенная грамматика абстрактной живописи оказывается пустой в том смысле, что выполнение этих правил может быть не связано с какой-либо специфической информацией. По-видимому, здесь полностью сохраняется свобода индивидуального выбора.

Итак, если мы будем рассматривать абстрактную живопись как систему знаков, то эта система оказывается вырожденной - здесь пустыми оказываются как первичная знаковая система, так и правила оперирования с этими знаками. Естественно поставить вопрос - можно ли такую систему назвать языком? По-видимому, у нас нет достаточных оснований не признавать языками вырожденные знаковые системы такого типа. В них в крайней своей форме проявляется та тенденция, которая проявляется уже и в обыденных, естественных языках, и отличает эти языки от строго формализованных искусственных языков. Кстати, заметим здесь, что если бы пустая знаковая система абстрактной живописи имела строгую грамматику, то такой язык был бы попросту графической формализацией какой-либо дедуктивной логической системы. Схему доказательства какой-нибудь даже совсем сложной теоремы вполне можно зарисовать, используя абстрактные знаки.

Попробуем показать, что знаковая система абстрактной живописи имеет все же какие-то общие черты с невырожденными системами, традиционно воспринимаемыми как языковые системы. Для этого нам надо теперь рассмотреть такие характеристики языка, как семиотика - взаимодействие с "передатчиком", прагматика - взаимодействие с "приемником", и,что особенно важно - возможность построения иерархических знаковых систем, стоящих над исходными знаковыми системами - алфавитом первичных знаков и их объединениями-картинами.

О взаимодействии с "передатчиком" мы мало что можем сказать, поскольку мы не знаем, что собственно хотел сказать художник в той или иной картине. Существует большая литература, пытающаяся дать те или иные обоснования творческого процесса при создании произведений абстрактной живописи. Одно из возможных объяснений - стремление к выражению подсознательного мира художника в том смысле, как это понимается в концепциях, связанных с З. Фрейдом. Мы не будем здесь рассматривать все эти вопросы - они выходят за рамки поставленной нами задачи. Для нас важно обратить внимание на то обстоятельство, что каждый выдающийся художник-абстракционист имеет свой стиль. Это значит, что он отдает предпочтение некоторым знакам первичного алфавита и некоторым правилам обращения с этими знаками. Даже совсем малоподготовленный зритель всегда легко опознает, скажем, картины Кандинского. Отсюда следует, что вполне осмысленной оказывается попытка связать стиль художника с особенностями его сознания или скорее - подсознания. Может быть, практически эта задача оказывается необычайно трудной, но с логических позиций ее постановка правомерна и, следовательно, семантический анализ языка абстрактной живописи, по крайней мере, принципиально, возможен.

Теперь перейдем к прагматическому анализу. Чтобы осуществить его, надо суметь предложить зрителям некоторую систему чтения. Один из возможных способов чтения - это линейное ранжирование по степени предпочтения, ранее подробно исследованное в [8]. Обратим внимание на то, что здесь при анализе языка абстрактной живописи мы впервые применяем некоторый достаточно формализованный прием. Правда, это не единственный возможный способ чтения - можно, например, предложить нелинейное ранжи-

рование, ранжирование с группированием и т.д. Но мы имеем право изучать любой способ чтения, если он даже в каком-то смысле будет не самым лучшим. Чтение картин можно рассматривать как некоторый процесс взаимодействия между зрителями-экспертами и картинами. Если, обрабатывая статистические результаты такого исследования, мы сможем разбить как зрителей, так и картины на некоторые группы, т.е. если покажем, что изученное взаимодействие не создает некоторого равномерного шумового поля, то у нас будут основания считать, что абстрактная живопись читается и, следовательно, мы докажем, что прагматический подход к этой знаковой системе возможен. Нам могут возразить, что результаты такого чтения неустойчивы - они могут зависеть от эмоциональной настроенности экспертов, от уровня их подготовки к восприятию абстрактной живописи. Но все эти возражения в одинаковой мере относятся и к обыденному языку. Если мы примем гипотезу о том, что слово можно рассматривать, как некий "черпак", зачерпывающий что-то в нашем сознании, то тогда станет ясным, что чтение обыденного слова также не очень ух устойчивый процесс. Он также может зависеть в какой-то степени и от эмоциональной настроенности и в очень большой степени от предварительной подготовки.

Теперь, наконец, нам осталось обсудить последнее, и, может быть, самое важное свойство языка - возможность представления его в знаковых системах разного уровня. Два уровня для языка абстрактной живописи представляются очевидными - это уровень алфавита первичных знаков (см. приложение I) и уровень собственно картин - каждая картина может рассматриваться как фраза в этом языке. Возможность построения некоторых классификаций картин при их чтении группой экспертов создает третий уровень.

Рассмотрим еще один способ построения высшего уровня знаковой системы. Мы можем представить себе многомерное пространство, координатами которого будут элементарные знаки картин и правила оперирования с этими знаками. Каждая картина по любой координате может принимать только два значения: + 1 , если данный признак присутствует, и 0, если он отсутствует. Разместив таким образом картины в нашем пространстве, можно, пользуясь некоторыми формальными методами, найти их скопления и получить знаковую систему более высокого уровня - систему, по отношению к которой элементарными знаками будут картины.

## 2. Статистический анализ данных эксперимента по восприятию произведений абстрактной живописи.

Для проверки изложенной выше концепции был проведен следующий эксперимент. 100 человек, которых мы называем экспертами, расставили в порядке предпочтения 19 репродукций картин, относящихся к различным направлениям абстрактной живописи (см. приложение 2). Кроме того, каждая картина была охарактеризована набором 36 языковых признаков, перечисленных в приложении 1. Таким образом, исходным материалом для статистического анализа были следующие матрицы: матрица 100 x 19 результатов ранжирований картин, и матрица 19 x 36 признаков картин.

Были поставлены следующие задачи.

Во-первых, требовалось найти группы, состоящие из "похожих" картин, т.е. произвести естественную классификацию картин, основанную на восприятии их экспертами.

Во-вторых, интересно было понять, что представляют собой полученные группы картин. С одной стороны, это делалось путем содержательного анализа картин и их распределения по группам.

С другой стороны, была сделана попытка описать найденные группы картин с помощью формального языка 36 признаков. Это дало возможность понять, как чтение картин, механизм которого нам неизвестен, связано с составленными нами алфавитом и грамматикой.

Группировка картин производилась с помощью одного из алгоритмов кластер-анализа (от английского "cluster" - гроздь), состоящего в следующем [9]. Вычисляется матрица корреляций между группируемыми элементами. Два наиболее тесно коррелированных элемента $x$ и $y$ объединяются в одну группу $x+y$, которая далее рассматривается как один элемент, причем корреляция между $x+y$ и некоторым другим элементом $z$ определяется как среднее арифметическое корреляций между $z$ и $x$ и между $z$ и $y$. Производя соответствующий пересчет, мы получаем матрицу корреляций на единицу меньшего порядка. Снова ищутся два наиболее близких элемента, они объединяются, и порядок матрицы корреляций понижается еще на единицу. Число групп также на каждом шаге уменьшается на единицу, и работа алгоритма прекращается, когда все элементы объединятся в одну группу. Если объединяющиеся элементы $u$ и $v$ соответствуют группам из $p$ и $q$ исходных элементов, то корреляция между $u+v$ и некоторым элементом $w$. вычисляется как среднее взвешенное корреляций $w$ с $u$ и $v$

$$r_{u+v,w} = \frac{p}{p+q} r_{u,w} + \frac{q}{p+q} r_{v,w}$$

На рис. I представлены результаты, полученные на последних шагах работы алгоритма. Удобнее рассматривать их в обратном порядке, т.е. рассуждать не в терминах объединения, а в терминах разделения. Если придерживаться этого соглашения, то процесс

описывается следующим образом. Исходная коллекция из I9 картин разделяется на две группы из 9 и из I0 картин. Каждая из этих групп в свою очередь делится затем на две группы, после чего делится на группы одна из получившихся четырех групп. Мы остановимся на этом разбиении и перенумеруем соответствующие группы картин, как это сделано на рис. I.

Наряду с группировкой с помощью указанного алгоритма, методом главных компонент [10] I00-мерное пространство оценок картин было спроектировано на плоскость. На рис. 2 показано расположение картин на этой плоскости. В принципе изображение на плоскости позволяет делать группировку просто на глаз. Однако, учитывая искажения, возникающие при понижении размерности, мы предпочли группировку, полученную с помощью кластер-анализа, которая также изображена на рис. 2. Тем не менее, наглядное представление картин на плоскости главных компонент оказалось очень полезным. В частности, стало видно, что наряду с объединением групп I, П, Ш, IУ и У в группы I + П и Ш + IУ +У, полученные с помощью кластер-анализа (рис. I), довольно естественным кажутся объединения I + Ш и IУ + У + П.

Содержательный анализ группировки картин проводился после того, как они были расположены в соответствии с рис. 2. Во-первых, следует отметить, что полученное разбиение очень хорошо соответствует зрительному впечатлению, и у нас не появлялось желания переместить какую-либо картину в другую группу. Наиболее легко видна разница между двумя крупными группами I + П и Ш + IУ + У. Если первая из них составлена из картин, представляющих собой довольно сложные композиции фигур в основном неправильной формы и отличающихся богатством цветов, то картины второй группы являются сравнительно простыми композициями пра-

вильных или почти правильных геометрических фигур. Различия внутри групп также прослеживались без особого труда. Так, если в группе I ощущалось наличие композиционного замысла и картины требовали внимательного изучения, то в картинах группы II искусственно создавалась некоторая хаотичность в форме и взаимном расположении элементов, и они явно аппелировали к непосредственному зрительному восприятию. Картины групп IУ и У отличаются от картин группы Ш ярко выраженной рельефностью - все они представляют изображения трехмерных фигур. Если, однако, пытаться найти различие, аналогичное различию между группами I и II, то, по-видимому, здесь также выделяется явная аппеляция картин групп IУ и У к непосредственному зрительному восприятию, но в отличие от картин группы II, в которых такой эффект достигался нарочитой небрежностью и искусственной хаотичностью, здесь используется предельная лаконичность.

Другой подход к анализу языка абстрактной живописи заключался в использовании формального языка 36 признаков. Чтобы связать эти признаки с найденными группами картин, к ним были добавлены еще 5 признаков принадлежности картин к группам I,II, Ш,IУ,У. Эти признаки принимали значение I, если картина входила в группу, номер которой совпадал с номером признака и 0 - в противном случае. После этого была вычислена матрица коэффициентов корреляции между полученными 4I признаками. Ее непосредственное рассмотрение дает возможность выяснить, какие признаки наиболее характерны для разных групп картин. Однако результаты становятся более наглядными, если, исходя из полученной матрицы корреляций, произвести кластер-анализ данных 4I признаков. Результаты последних шагов работы алгоритма кластер-анализа приведены на рис. 3.

Первое, что бросается в глаза, это то что к четвертой группе не присоединился ни один из признаков. Так как картины этой группы характеризуются ярко выраженной рельефностью, то, по-видимому, отсутствие признаков, характеризующих объемность изображения, следует отнести к недостаткам составленного нами алфавита и грамматики. Что касается остальных групп, то с каждой из них ассоциируется несколько признаков - с первой группой - I3, со второй - 8, с третьей - 4, и, наконец, с пятой - 2 признака. Эти цифры дают некоторое представление о сложности картин, входящих в каждую из групп, что вполне согласуется с непосредственным зрительным впечатлением.

На основании этих результатов мы можем найти место групп картин в нашей языковой концепции. Очевидно, картины внутри каждой из групп находятся в парадигматическом отношении - отношении, которое возникло у экспертов в процессе чтения картин из-за ассоциативных связей, вызванных сходством алфавита и грамматических правил, используемых в картинах одной группы (см. [14] , стр. 432). А тогда эти группы картин можно назвать парадигмами. Очевидно, мы при этом несколько расширяем содержание этого термина по сравнению с его обычным употреблением при описании языковых конструкций, но о такой возможности говорилось еще в самом начале этой работы.

Попробуем теперь рассмотреть различные объединения парадигм I, II, III, IУ и У, что определяется поставленной нами задачей получения различных уровней иерархии в изучаемой нами языковой системе.

Вернемся к полученным ранее объединениям I + II и III + IУ+У и I + III и II + IУ + У. Назовем первую пару парадигм парадигмами A и B, а вторую пару - парадигмами C и D . Изложенным не-

много выше приемом исследуем различия между этими объединениями. Результаты кластер-анализа 38 признаков (36 признаков языка и признаки *A* и *B* ) даны на рис. 4. Картина получилась довольно неожиданная. Больше половины признаков вообще не присоединилось ни к одной из парадигм А и В. Что касается остальных, то за исключением одного признака (симметрия), все они ассоциировались с парадигмой А. Это говорит, в сущности, о том, что объединение картин В не является парадигмой, так как входящие в нее картины слишком разнородны по используемому алфавиту и грамматике. Если рассмотреть теперь таким образом парадигмы С и *D* (рис. 5), то мы увидим, что все признаки с самого начала разбиваются на две части, одна из которых связана с парадигмой С, а другая с парадигмой *D* . Парадигмы I + Ш и П + IУ + У, следовательно, являются более осмысленными по нашему определению, чем парадигмы I + П и Ш + IУ + У, и представляют собой высший уровень иерархии в нашей системе.

Исходя из этих результатов, интересно было выяснить, также как оценивают эксперты полученные парадигмы. Поскольку следовало ожидать,что их мнение будет различным, то прежде всего были найдены группы, состоящие из экспертов, сходным образом ранжирующих картины. С помощью описанного выше алгоритма кластер-анализа были выявлены три группы, состоящие из 79, IO и II экспертов. Для каждой из этих групп были вычислены средние ранги парадигм I, П, Ш и IУ + У.

Из таблицы I видно, что первая группа экспертов высоко оценивает парадигмы I и П, т.е. парадигму А, вторая группа предпочитает парадигмы I и Ш, т.е. парадигму С,и третья - парадигму IУ + У.

Была сделана попытка связать эти различия в восприятии

картин с социально-демографическими характеристиками экспертов. С этой целью им было предложено ответить на ряд вопросов, например, таких как пол, возраст, образование родителей, род занятий, степень знакомства с абстрактной живописью и отношением к ней и т.д. - всего 33 вопроса с ответами "да" и "нет". К характеристикам экспертов по этим вопросам были добавлены еще 3 признака, указывающих на принадлежность каждого эксперта к одной из трех выделенных нами групп, после чего была вычислена матрица парных корреляций, и по ней, как и ранее, проведен кластер-анализ, результаты которого представлены на рис. 6.

Прежде всего мы видим, что ряд характеристик не ассоциируется ни с одним из трех типов восприятия. Что касается остальных, то они распределяются по группам следующим образом. Для первой группы характерны отрицательные ответы на вопросы о посещении художественных музеев и временных художественных выставок и о покупке книг по искусству, незнание творчества Кандинского, наличие только лишь начального образования у отца и матери и работа в области технических наук. Вторая группа состоит в основном из представительниц женского пола, с возрастом до 30 лет, рутинным характером труда. Эта группа систематически покупает книги по искусству, посещает временные художественные выставки.

В третью группу входят эксперты, имеющие гуманитарное образование с априорным отрицательным отношением к абстрактной живописи и творчеству Кандинского, при осмотре картины обращающие в первую очередь свое внимание на ее название.

Из полученных ранее выводов о содержании картин, входящих в парадигму А видно, что высокая оценка первой группой экспертов именно этой группы картин объясняется не только их склон-

ностью к положительной оценке тех или иных особенностей языка абстрактной живописи, сколько, возможно, просто подсознательным уважением к трудности этих картин.

Что касается второй группы экспертов, то хорошая осведомленность экспертов в вопросах искусства, по-видимому, и послужила основой высокой оценки картин парадигмы С, включающей картины известных художников-абстракционистов - В. Кандинского и К. Малевича - и отличающихся большим единством языковых признаков, нежели парадигма А. Для третьей группы экспертов, состоящей из противников абстрактной живописи, кажется естественным то, что из всех картин они отдали предпочтение изображениям трехмерных фигур.

3. Заключение.

Итак, мы видим, что зрители могут читать картины, связывая их во все более и более сложные объединения - парадигмы. Выполняется требование иерархичности языка. Люди могут быть разбиты на группы по их умению строить парадигмы одинаковым образом. Отсюда, а также из предыдущих рассуждений следует, что абстрактную живопись можно рассматривать как язык. Правда, это совсем особый - вырожденный язык, и как раз поэтому он представляет большой интерес для изучения.

Хочется обратить внимание на то, что обычная предметная живопись не может рассматриваться аналогичным образом. Для нее трудно построить алфавит первичных знаков. Это будет просто набор мазков или штрихов разной формы. Вряд ли представляет какой-либо интерес рассматривать такой примитивный алфавит. Совсем уже трудно составить грамматику, задающую правила упорядочения мазков. Видимо, элементарной единицей предметной живописи надо считать образ. По-видимому, образ может превращаться в

знак, если им можно оперировать по некоторому произволу, разрушая принятый стереотип. Скажем, женская грудь в живописи есть образ, но если, следуя сюрреалистическим приемам, ее можно поместить на любое место человеческого тела, то это уже будет знак.

Можно говорить, по крайней мере, о трех уровнях интеллектуальной деятельности людей: об образном - дологическом уровне, о логическом уровне и возможно - о непонятом еще, сверхлогическом уровне, на котором, по-видимому, протекают творческие процессы. Все три уровня могут быть присущи одному и тому же человеку. Конкретная живопись, создающая образы, есть средство коммуникации, соответствующее образному мышлению. Язык как система знаков и правил оперирования с ними есть средство коммуникации на логическом уровне. Правда, для языка есть широкий диапазон возможностей - его структура может быть строго формализованной, в соответствии с требованиями дедуктивной логики,или быть слегка разболтанной полиморфизмом языка и нечеткостью грамматики, или, как это имеет место в абстрактной живописи - может быть совершенно свободной. Абстрактная живопись есть попытка создания искусства, использующего средства коммуникации логического уровня. Естественно, что это удалось сделать,только создав особый - вырожденный язык. Если бы язык оставался невырожденным и сохранял известный формализм, все это было бы лишь некоторой пародией на логику.

Имеет ли право на существование такой вырожденный язык? Есть ли это действительно искусство? На эти вопросы наша работа ответа не дает. Абстрактная живопись как некий феномен существует, независимо от того, нравится это нам или нет. И этот феномен может подвергаться изучению формальными методами. Ре-

зультаты такого исследования прежде всего обогащают наше представление о природе языка. Мы видим, что структура языка как знаковой системы может принимать две крайних формы - быть строго дедуктивной системой, или быть системой современно свободной - тогда она интерпретируется как одна из форм искусства. Обыденный язык людей занимает промежуточное положение. В отдельных случаях, в зависимости от об области применения, он может преимущественно приближаться то к одной, то к другой своей грани.

Основываясь на проделанной нами работе можно сформулировать следующие соображения, относящиеся к изучению знаковых систем:

I. При семантическом анализе знаковой системы надо обращать внимание на возможность осмысленного (не обязательно адекватного) перевода содержания текста на язык другой знаковой системы. Можно даже высказать такое утверждение - знаковые системы можно рассматривать как языковые только тогда, когда возможно на одной из них что-то сказать о текстах, написанных на другой, хотя бы аналогично тому, как мы сделали в этой работе, рассказав на другом - обыденном-языке о парадигмах абстрактной живописи. Будем считать, что в этом случае мы сумели сделать какой-то, хотя бы и неадекватный перевод с одного языка на другой.

2. При семантически-прагматическом анализе знаковой системы можно столкнуться с тем, что "приемник" будет воспринимать совсем не то содержание, которое было вложено в сообщение передатчиком. Для пояснения этой ситуации можно попытаться построить вероятностную модель прагматики, используя хоро-

но известную в математической статистике теорему Бейеса[10]. Поясним смысл этой теоремы в обычных статистических терминах. Допустим, что производится измерение величины $\mu$ для некоторого объекта $H$. Имеется пространство $Y$ всех возможных результатов измерений $y$. На этом пространстве задана вероятность $P(y/\mu)$, в простейшем случае это просто функция нормального распределения для ошибок наблюдений при измерении объекта $H$. Далее, будем считать, что нам известна априорная вероятность $P(\mu)$, т.е. априори ( до проведения опыта) известно распределение всех возможных значений $\mu$. Тогда теорему Бейеса можно записать так

$$P(\mu/y) = k\,P(y/\mu)\,P(\mu)$$

где $k$ - константа, полученная из условия нормировки. Вводя в рассмотрение априорную информацию, мы как бы задаем вход в систему, а затем, используя теорему Бейеса, образуем логически безупречным образом выход из системы, который записывается в виде апостериорной вероятности $P(\mu/y)$. Если теперь мы вернемся к анализу процесса восприятия знаковой системы, то в этом случае $P(\mu)$ будет априорным распределением суждений о смысловом значении знака. Это распределение может быть построено, скажем, так: "приемник" имеет в своем сознании некоторое представление о возможных смысловых значениях знака - одно из них имеет большую вероятность появления, другое меньшую и т.д. Все это может быть представлено функцией

---

[10] Напомним здесь, что так называемый необейсовский подход сейчас широко используется для логического обоснования построения суждений в математической статистике, см., например, [12] .

распределения, построенной так, что по оси абсцисс отложены ранги смысловых значений, установленных по вероятности их появления, по оси ординат отложены сами вероятности. Эта априорная вероятность создаст вход в систему восприятия читаемого текста. Система чтения позволяет образовать функцию $P(y/\mu)$ - она задается многими факторами - способом комбинирования читаемого знака с другими знаками фразы и общей эмоционально-интеллектуальной настроенностью "приемника" в момент чтения; последнее обстоятельство вносит тот же элемент неопределенности, что и ошибка эксперимента в обычных физических экспериментах. В частном случае полного априорного незнания (или априорного безразличия) функция $P(\mu)$ будет просто равномерным распределением и тогда $P(\mu/y)$ сведется к $P(y/\mu)$ ,но вряд ли это может иметь место, когда "приемником" является человек. Если $P(\mu)$ для "приемника" и "передатчика" более или менее одинаковы, то процесс чтения будет вносить только случайные искажения. Но может оказаться , что "приемник" и "передатчик" вкладывают совершенно разный смысл в знаковую систему, как это, например, имеет место при эстетическом восприятии коллекции насекомых. В этом случае у "приемника" - человека имеется некоторая система эстетических представлений, задаваемая функцией $P(\mu)$ , которой нет у "передатчика", т.е. у системы генетической информации насекомых. Отсюда и возможность восприятия того сообщения, которое не высказывалось "передатчиком". Если хотите, рассмотренная здесь модель - это есть просто перевод на вероятностный язык приведенного выше упрощенного высказывания о том, что восприятие слова происходит по схеме черпака, зачерпывающего из созна-

ния человека то, что там имеется. Вероятностный процесс восприятия текста особенно легко наблюдать при чтении иероглифических языков.

Рассматриваемая здесь вероятностная модель практически, конечно, нуждается в более глубокой разработке. Вряд ли она может служить рабочим инструментом для каких-либо конкретных вычислений. Но она, как нам кажется, может помочь понять те трудности, с которыми пришлось столкнуться при попытке использовать формальную логику для отождествления знаковых структур текстов и анализа заложенного в них содержания. Прекрасное изложение логической семантики, данное в [13] , заканчивается грустным замечанием о "несбывшихся надеждах", возлагавшихся в информатике на хорошо известные методы логики. Почему эти, столь радужные, в недавнем прошлом, надежды не сбылись? Ответ здесь простой - восприятие текста это вероятностный процесс, лингвистическая семантика невыразима через логическую семантику.

Заканчивая эту работу, нам хочется ответить на два возможных возражения. Первое из них сводится к тому, что формально такой же результат мы могли бы получить и для какой-либо другой коллекции зрительных образов, скажем для коллекции бабочек. Можно ли утверждать, что форма насекомых, их вид есть некоторый язык? На первый взгляд на этот вопрос надо дать отрицательный ответ - здесь нет процесса коммуникации, того процесса, который собственно и создает язык. В противном случае мы должны были бы признать, что есть природа, как некоторое существо, ведущее разговор с человеком. Но исходя из развитой выше концепции прагматики с привлече-

нием Бейесовского подхода, здесь возможна и другая интерпретация. Мы можем считать, что популяция насекомых ведет разговор с окружающей ее природой, предлагая для процесса эволюции все время изменяющееся многообразие признаков. Люди, сумев проранжировать коллекцию насекомых, прочли эту запись совсем с других позиций - исходя из своих априори заданных эстетических представлений. Не возникнет ли похожая ситуация, когда придется читать сообщения жителей других миров?

Второе возражение - не слишком ли мы упрощаем задачу творчества, записав содержание картин с помощью алфавита и грамматики? Не значит ли это, что пользуясь этим алфавитом и грамматикой картину может воспроизвести ЭВМ? Действительно, зная алфавит и грамматику, машина, вообще говоря, должна суметь воспроизвести картину. Но в нашем случае язык вырожденный, и картина, воспроизведенная машиной, будет находиться среди очень большого множества других похожих картин. Остается неясным, как именно эту картину, среди других однотипных, отобрал художник в процессе своего творчества.

Такой же вопрос у нас возникает, когда мы пытаемся моделировать на ЭВМ творчество композитора, или когда мы строим модель творческого мышления, вводя генератор случая для нарушения строго дедуктивных форм мышления у творчески активных ученых, создающих новые концепции в науке. Всегда остается без ответа один и тот же вопрос: как же устроен алгоритм для отбора действительно талантливого решения. А не зная ответа на этот вопрос мы не можем говорить о том, что моделируем на ЭВМ творческий процесс.

Нам хочется надеяться, что исследования подобного рода обогащают наше представление о языке как о кибернетической категории. А если все изложенное выше кто-то воспримет просто как некую спекуляцию, то мы ответим следующее : это все же изящная спекуляция, достойная обсуждения [11] .

---

11 После того, как эта работа была закончена, мы познакомились с рефератом статьи Линдауера [11] , в которой изучался вопрос о предпочтительной ориентации в абстрактной живописи двух групп экспертов: художников-профессионалов и студентов-нехудожников. К сожалению, полный текст этой статьи оказался для нас недоступным.

Литература.


I. Д.П. Горский - От описательной семиотики к семиотике теоретической. Журнал "Вопросы философии", № IO, I969,стр. 72-8I.

2. И.А. Мельчук, Р.М. Фрумкина - Кибернетика и некоторые проблемы современной лингвистики. Статья в сб. "Кибернетика на службу коммунизму", ред. акад. А.И. Берг, т. 3, 294-302, изд-во "Энергия", I966.

3. Ю.А. Шрейдер - К вопросу об определении основных понятий семиотики. Статья в том же сборнике, стр. 26I-274.

4. C. Laird - Thinking about Language, Half, Reinhard and Winston, N. Y., I96I.

5. Э. Геллнер - Слова и вещи, пер. с англ. МЛ, I962, 343 стр.

6. В.В. Налимов, З.М. Мульченко - Логико-лингвистический анализ языка науки (подготовлено к печати).

7. Э. Нейгел, Дж. Ньюмэн - Теорема Геделя, М., Знание, I970, 62 стр. Сокращенный перевод с английского.

8. П.Ф. Андрукович, Г.Н. Веселая, А.А. Каменский, В.Н. Кознрев, В.В. Налимов, А.Т. Терехин - Статистический анализ экспертных оценок. Труды НИКФИ, Симпозиум по субъективному анализу изображений. (В печати, I97I).

9. R. R. Sokal, D. C. Michener - A Statistical Method for Evaluating Systematic Relationships. Univ. Kansas Sci. Bull., I958, v. 38, I409-38.

IO. Т. Андерсон - Введение в многомерный статистический анализ. М., Физматгиз, I963.

II. M. S. Lindauer - The Operation Form in Abstract Art. Proc. of the 77-th Ann. Conv. of the Amer. Psych. Ass., I969, part I.

I2. В.В. Налимов - Теория эксперимента, "Наука" (в печати).

I3. В.Г. Влэдуц, В.А. Успенский, Ю.А. Шрейдер - Семантика и научная информация (доклад на Международном симпозиуме стран членов СЭВ, Москва, 9-I3 июня I970 г.), "Теоретические основы информатики", изд. ВИНИТИ.

I4. А.И. Михайлов, А.И. Черный, Р.С. Гиляревский, Основы информатики, М., "Наука", I968.

Приложение I.

Алфавит абстрактных картин.

I) Линия.

2) Полоса.

3) Правильная геометрическая фигура (круг, треугольник, квадрат, прямоугольник, трапеция), выпуклая.

4) Фон.

5) Цвет.

6) Диффузные облака.

Грамматика абстрактных картин.

I. Операция над линиями:

I) прерывание,

2) изгиб,

3) излом,

4) пересечение (линия пересекает другие линии),

5) периодичность линий.

II. Операции над полосами:

6) прерывание,

7) изгиб,

8) излом,

9) пересечение (полоса пересекает себя или другие полосы с наложением или с проникновением),

10) обрамление полосы цветом (тоном),

11) периодичность полос.

III. Операции над фигурами:

12) дробление резанием,

13) дробление цветом,

14) деформация (сжатие, растяжение, поворот);

15) наложение фигур,

16) пересечение фигур,

17) проникновение - переплетение фигур,

18) обрамление фигуры цветом (тоном),

19) периодичность фигур.

IУ. Операция над фоном:

20) дробление,

21) плавное изменение.

У. Операции над цветом:

22) наличие чистых цветов,

23) заполнение фигуры (полосы) одним цветом,

24) заполнение фигуры (полосы) несколькими цветами,

25) взаимопроникновение цветов,

26) периодичность цвета.

УI. Операции над диффузными облаками:

27) изменение цвета (тона) внутри облака,

28) взаимопроникновение облаков.

УП. Смешанные операции.

29) пересечение линией полосы,

30) пересечение линией геометрической фигуры,

31) пересечение полосой геометрической фигуры,

32) пересечение линией диффузной фигуры,

33) пересечение полосой диффузной фигуры.

УШ. Операции общего характера:

34) создание преимущественного линейного направления организации,

35) создание кругового направления организации,

36) создание симметрии.

Таблица I. Средние ранги парадигм A, B, C и D , соответствующие трем группам экспертов.

| | Первая группа | Вторая группа | Третья группа |
|---|---|---|---|
| A | 6,9 | 8,2 | II,6 |
| B | 7,9 | I3,8 | II,5 |
| C | I2,5 | 6,8 | 9,6 |
| D | I0,3 | I3,0 | 6,3 |

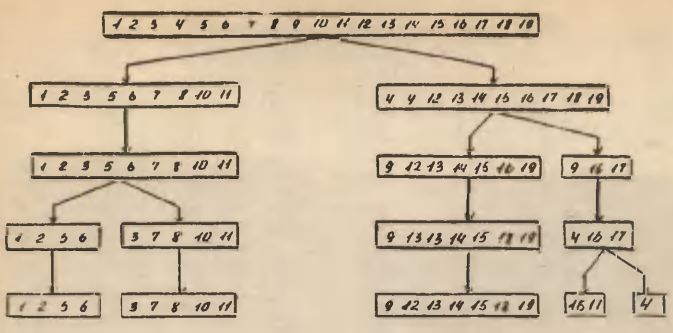


Рис. 1. Кластер-анализ 19 картин по оценкам 100 экспертов

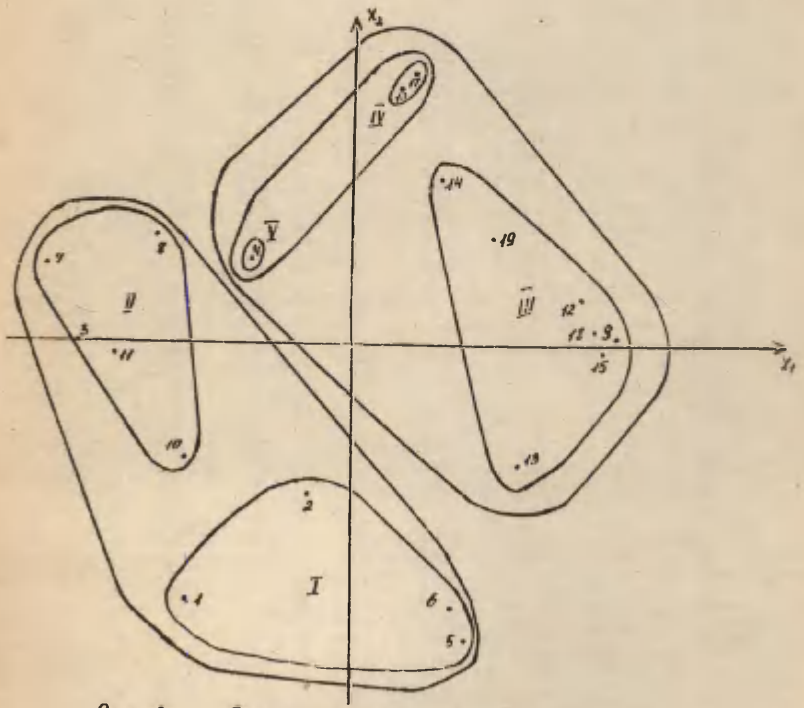



Рис. 2 Расположение картин в плоскости двух первых компонент $X_1, X_2$.

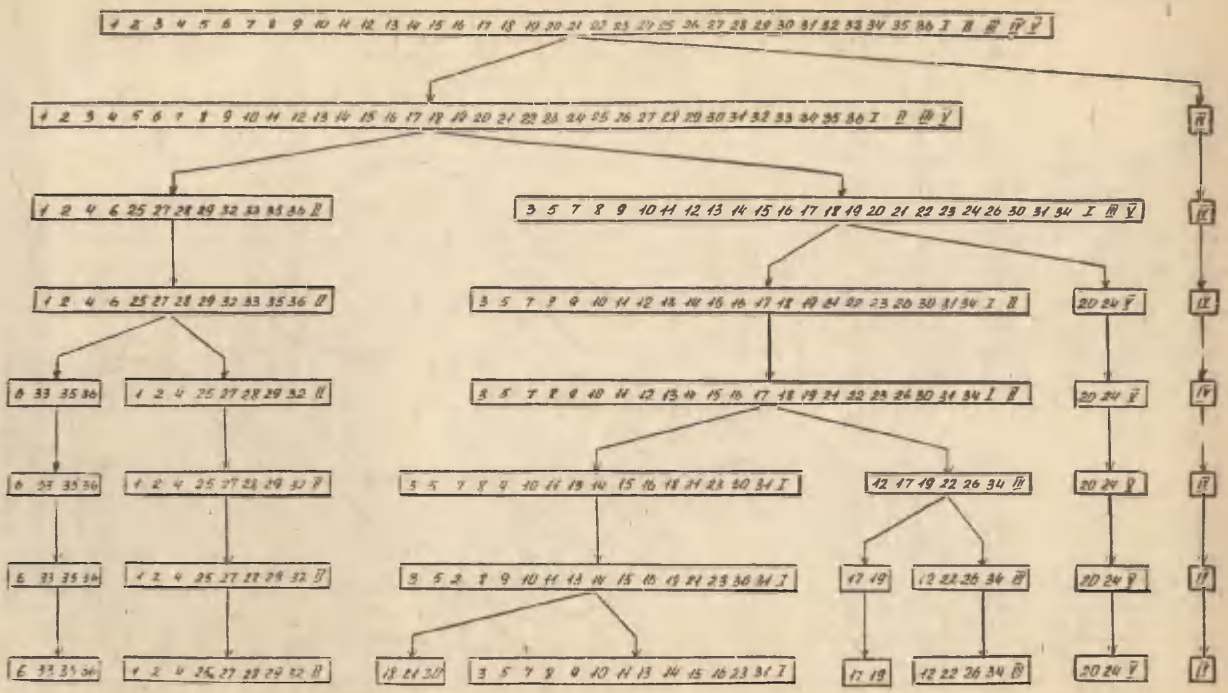


Рис 3 Кластер-анализ 41 признака (36 языковых признаков дополнены признаками I-V)

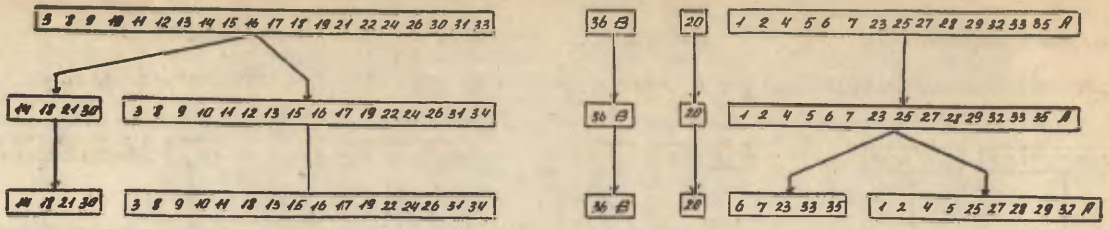


Рис.4. Кластер-анализ 38 признаков (36 языковых признаков дополнены признаками A и B)

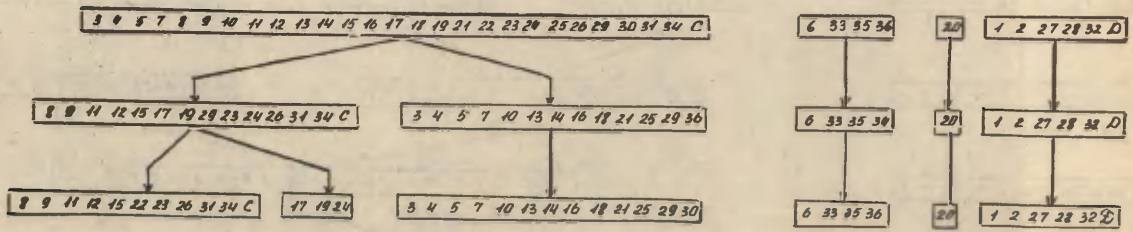


Рис 5. Кластер-анализ 38 признаков (36 языковых признаков дополнены признаками C и D)

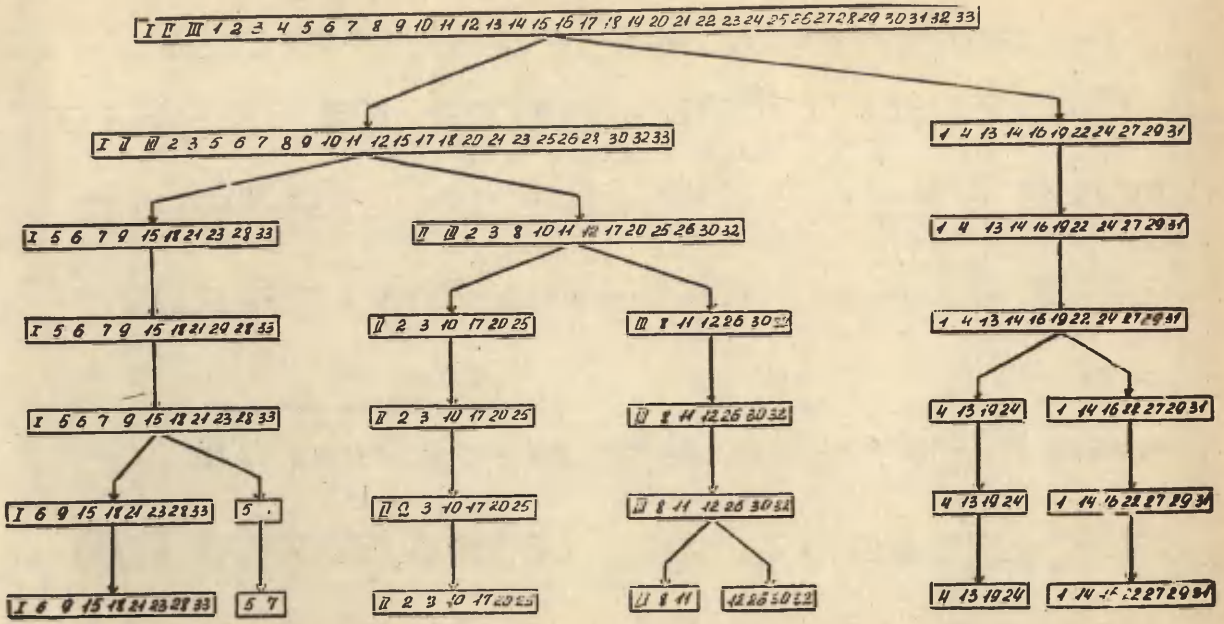


Рис.6 Кластер-анализ 36 характеристик (33 социально-демографические характеристики дополнены индексами принадлежности к одному из трех типов восприятия.)

# ABSTRAKTNE KUNST KUI OMAPÄRANE KÕDUNUD KEEL (STATISTILISED MEETODID ABSTRAKTSE KUNSTI TAJUMISE UURIMISEKS)

P.F. Andrukovitš , V.S. Gribkov, V.P. Kozõrev ,
V.V. Nalimov, A. T. Terjohhin
(Moskva)

## Resümee

Käesoleva srtikli eesmärgiks on loogilise analüüsi alusel kontrollida võimalust interpreteerida abstraktset kunsti märgisüsteemina, mis esindab omapärast keelt.

Keelt mõistavad autorid informatsiooni edasiandmise süsteemina, mille aluseks on esmane märkidesüsteem.

Esitatakse (lisa I) abstraktse kunsti aluseks olev esmane märkide süsteem. Märgitakse, et keele struktuuri kohaselt ei ole kodeerimise ja dekodeerimise ühesuse nõue oluline.

Abstraktse kunsti esmases tähesüsteemis puuduvad siduvad reeglid, s elletõttu on õige seda keelt nimetada kõdunuks. Ka abstraktse kunsti grammatika, mis on fenomenoloogiliselt koostatud (vt. lisa I) osutub tühjaks: vahekorrad ei ole seotud mingi informatsiooniga.

Autorid püüavad esmakordselt rakendada formaliseeritud metoodikat abstraktse kunsti käsitlemiseks.

Ekspertide rühmal (100 inimest) lasti järjestada 19 kunstiteose reproduktsioonid, igaüks neist oli 36 keeletunnusega iseloomustatud.

Ekspertidelt nõuti ka piltide rühmitamist. Seejärel kasutati ka formaalselt rühmitamise algoritmi, mille rakendamise tulemusi esitab joonis 1.

Analüüsitakse erinevate tunnuste alusel saadud rühmitusi sisuliselt.

# ABSTRACT ART AS A PECULIAR DEGENERATE LANGUAGE (STATISTICAL METHODS APPLIED IN INVESTIGATING ABSTRACT ART PERCEPTION)

P.F. Andrukovitch, V.S. Gribkov, V.P. Kozyrev, V.V.Nalimov, A.T. Teryokhin

(Moscow)

Summary

The aim of the present article is to test on logical grounds the possibilities of interpreting abstract art as a system of signs representing a peculiar language.

Language is conceived of as a system of information transmission, a primary sign system being the basis of it.

The primary sign system on which abstract art is based is presented. The uniqueness of coding and decoding is not important proceeding from language structure.

In the primary sign system of abstract art there are no linking rules therefore the language is degenerate. The grammar of abstract art constructed phenomenalistically is devoid: relations contain no information.

A formalized methodics is applied in treating the problem. The experts (100 judges) were made to arrange in order the reproductions of 19 pieces of art each of which was characterized by 36 language variables.

The grouping of pictures was also demanded. After that a formalized grouping algorithm was also used.

The groupings based on several variables are analysed.

## СПИСОК ДОКЛАДОВ,

прочитанных на семинаре и не опубликованных в настоящем сборнике

А.Г. И в а х н е н к о (Киев). Многорядная теория решения.

Ю.И. Н е й м а р к (Горький). Нормирование признаков и преобразование пространства признаков.

А.Г. Ф р а н ц у з (Ленинград). Некоторые вопросы оптимизации процедур обучения распознавания образов.

Л.К. В ы х а н д у (Таллин). Об опыте работы с большими массивами.

В.К. М а л у ш е н к о (Киев). Распознавание образов непрерывной речи.

И.Ю. И с т о ш и н (Новосибирск). Задача определения устойчивости типов рабочих по времени.

Ю. В о о г л а й д, А. М у р у т а р (Тарту). Опыт лаборатории социологии ТГУ при применении классификации в социологии.

СОДЕРЖАНИЕ

РАСПОЗНАВАНИЕ ОБРАЗОВ

(Материалы конференции)

На русском языке

Тартуский государственный университет
ЭССР, г.Тарту, ул.Юликооли,I8

Ответственный редактор Э.Тийт

Ротапринт ТГУ I972. Подписано к печати I6/У-I972 г.
Печ.листов I5,63 (условных I4,44). Учетн.-издат.
листов 9,93. Тираж 500 экз. Бумага 30х42.I/4.
МВ I2248. Заказ № 608.

Цена 70 коп.

Цена 70 коп.