UNIVERSITY OF TARTU
Institute of Computer Science
Software Engineering Curriculum

Artem Mateush

Automated Payment Classification in Retail Banking

Master's Thesis (30 ECTS)

Supervisor: Rajesh Sharma

Tartu 2018

Automated Payment Classification in Retail Banking

Abstract:

Retail banks use special techniques to analise their customer data to achieve business goals or improve their service. Modern machine learning techniques can be utilised to augment the classic data analysis techniques in this field. The ability to classify payments of their customers enables retail banks to better understand their customers' expenditure patterns and to customize their offers accordingly. Payment classification is a difficult problem because of the large and evolving set of businesses and the fact that each business may offer multiple types of products, e.g. a business may sell both food and electronics. Two major approaches to payment classification are rule-based classification and machine learning-based classification. The classification machine learning technique is a variant of supervised learning, and, as such, it requires a labeled transaction set — in our case, transactions classified by the customers themselves (as a form of crowdsourcing). The rule-based approach is not scalable as it requires rules to be maintained for every business and type of transaction. The crowdsourcing approach leads to inconsistencies and is difficult to bootstrap since it requires a large number of customers to manually label their transactions for an extended period of time.

Here we present a case study at a financial institution in which a hybrid approach is employed. A set of rules is used to bootstrap a financial planner that allowed customers to view their transactions classified with respect to 66 categories, and to add labels to unclassified transactions or to re-label transactions. The crowdsourced labels, together with the initial rule set, are then used to train a machine learning model. We evaluated our model on real anonymised dataset, provided by the bank, which consists of wire transfers and card payments. In particular, for the wire transfer dataset, the hybrid approach increased the coverage of the rule-based system from 76.4% to 87.4% while replicating the crowdsourced labels with a mean AUC of 0.92, despite inconsistencies between crowdsourced labels.

This improvement shows the viability of hybrid models proposed, and the positive evaluation result allows us to set up the integration of the hybrid model with the bank's systems.

Keywords:
payment classification, machine learning, crowdsourced data

CERCS: P170, Computer science, numerical analysis, systems, control

Tüübituletus neljandat järku loogikavalemitele

Lühikokkuvõte:

Selleks, et saavutada oma ärilisi eesmärke ja parendada teenusepakkumist, kasutavad jaepangad spetsiaalseid tehnikaid oma klientide andmete analüüsimisel. Kaasaaegseid masinõppe tehnikaid saab selles valdkonnas kasutada täiendusena klassikalistele

andmeanalüüsi meetoditele. Oskus oma klientide makseid klassifitseerida võimaldab jaepankadel oma klientide kulutuste mustritest paremini aru saada ja oma pakkumisi spetsiaalselt kohandada. Maksete klassifitseerimine on raske probleem, kuna äriklientide hulk on suur ja muutuv ja kuna iga äriklient võib pakkuda mitut tüüpi tooteid, näiteks võib müüa nii toitu kui elektroonikat. Kaks maksete klassifitseerimise põhilist lähenemist on reeglitepõhine ja masinõppe põhine klassifitseerimine. Masinõppepõhine klassifitseerimismeetod on supervised õppe vorm, ja sellisena vajab ta märgendatud andmeühikute kogumit - meie puhul klientide endi poolt klassifitseeritud transaktsioone (mis on oma olemuselt crowdsourcing). Reeglitepõhine lähenemine ei ole skaleeruv, sest see vajab iga äri ja transaktsioonitüübi jaoks hallatavat reeglite kogumit. Crowdsourcing põhine lähenemine toob endaga kaasa vasturääkivused ja seda on alguses raske käivitada, kuna vajatakse suure hulga klientide poolt, pika ajaperioodi jooksul, manuaalselt märgendatud transaktsioonide kogumit. Siinkohal toome ära finantsasutuse kaasusuuringu, mille raames on kasutatud hüübriidlähenemist. Kasutusel on finantsplaneerimise tööriist, mille käivitamiseks on loodud esmane reeglite kogum, ja klientidele on selle raames loodud võimalus vaadelda oma transaktsioone klassifitseerituna 66 kategooriasse ning lisada märgendeid klassifitseerimata transaktsioonidele või uuesti märgendada juba märgendatud transaktsioone. Crowdsourcetud märgendeid ja algset reeglite kogumit kasutatakse seejärel masinõppe mudeli treenimisel. Me hindame oma mudeli tõhusust elust võetud anonümiseeritud andmestikku kasutades, mille oleme saanud pangalt. See koosneb kontomaksetest ja kaardimaksetest. Täpsustades võib öelda, et kontomaksete andmestikul parandas hübriidlähenemine reeglitepõhise süsteemiga võrreldes katvust 76.4%-lt 87.4%-le, mille juures crowdsourcinguga leitud märgendeid replitseeriti 0.92 keskmise AUC juures, ja seda olenemata crowdsourcetud märgendites leiduvatest vasturääkivustest. Selline süsteemi edasiarendus viitab väljapakutud hübriidmudeli põhjendatusele, ning positiivne hinnang tulemustele võimaldab meid seadistada ja integreerida hübriidmudeli panga süsteemidega.

Võtmesõnad:
makse klassifikatsioon, masin õppe, crowdsourced andmed

CERCS:P170, Arvutiteadus, arvutusmeetodid, süsteemid, juhtimine (automaatjuhtimisteooria)

Acknowledgments This work is supported by an unnamed financial institution and the Archimedes European Regional Development Funds. We thank the employees of the financial institution who volunteered to create the external validation dataset.

Appendix

I. Licence

Non-exclusive licence to reproduce thesis and make thesis public

I, **Artem Mateush**,

1. herewith grant the University of Tartu a free permit (non-exclusive licence) to reproduce, for the purpose of preservation and making available to the public, including for addition to the DSpace digital archives until expiry of the term of validity of the copyright of my thesis

   Automated Payment Classification in Retail Banking

   supervised by Rajesh Sharma

2. Making the thesis available to the public is not allowed.

3. I am aware of the fact that the author retains the right refered to in point 1.

4. This is to certify that granting the non-exclusive licence does not infringe the intellectual property rights or rights arising from the Personal Data Protection Act.

Tartu, 09.08.2018