UNIVERSITY OF TARTU
Institute of Computer Science
Computer Science Curriculum

Mariia Godgildieva

Evaluation Metrics for Predictive Monitoring Systems with Highly Imbalanced Datasets

Master's Thesis (30 ECTS)

Supervisor: Marlon Dumas

Supervisor: Pavlo Tertychnyi

Tartu 2020

Evaluation metrics for predictive monitoring systems with highly imbalanced datasets

Abstract: A predictive monitoring system is a machine learning model used periodically with the goal of monitoring the behaviour of database entities. Most monitoring systems are trained and tested on highly imbalanced data as the target events are quite rare. Moreover, the evaluation of predictive monitoring is even more complicated by aspects specific to the task (e.g. proper timing of alerts and possible reoccurrence of alerts). Thus, there is a need in stable, class imbalance tolerant metrics that also reflect all monitoring-specific issues.

We have investigated existing approaches of monitoring systems evaluation and found them to be quite case-specific. Therefore, we have extended and modified the methods in use to be domain-independent and easily adjustable to the task at hand. The proposed evaluation approach is implemented and evaluated with experiments on data from different domains. In addition, we analysed several metrics designed specifically for imbalanced data to conclude if they can be used for monitoring evaluation due to restrictions of the approach.

Keywords: Predictive monitoring, machine learning, evaluation metrics, class imbalance

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

Hindamismõõdikud ennustavatele seiresüsteemidele kasutades tasakaalustamata andmestikke

Lühikokkuvõte: Ennustav seiresüsteem on masinõppe mudel, mida kasutatakse perioodiliselt, eesmärgiga teostada seiret andmebaasi entiteetidel. Enamus seiresüsteemid on treenitud ja testitud tasakaalustamata andmestikel, kuna sihtsündmused on haruldased. Lisaks on ka seiresüsteemide hindamine keeruline ülesandele vastavate spetsiifiliste nõudmiste tõttu (nt. häirete õigeaegne tõstmine ja häirete kordumine). Seega on vaja stabiilset, tasakaalustamata klassidele tolerantset hindamismõõdikut, mis võtab arvesse ka seiresüsteemile spetsiifilisi nõudmisi. Me oleme uurinud olemasolevaid lahendusi seiresüsteemidele ja leidnud, et enamus on juhtumipõhised. Seega me oleme laiendanud ja muutnud olemasolevaid meetodeid, et need oleksid valdkonnast sõltumatud ja lihtsasti kohandatavad vastavalt ülesandele. Väljapakutud hindamismõõdikud on rakendatud ja hinnatud eri andmestikke peal, mis on pärit eri valdkondadest. Lisaks oleme analüüsinud mitmeid mõõdikuid, mis on spetsiaalselt väljatöötatud tasakaalustamata andmestikke jaoks, et kindlaks teha, kas neid on võimalik kasutada piiratud lahendustes.

Võtmesõnad: Ennustav seiresüsteem, masinõpe, hindamismõõdikud, tasakaalustamata andmestikud

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

Appendix

I. Licence

Non-exclusive licence to reproduce thesis

I, **Mariia Godgildieva**,

1. herewith grant the University of Tartu a free permit (non-exclusive licence) to

   reproduce, for the purpose of preservation, including for the purpose of preservation in the DSpace digital archives until the expiry of the term of copyright,

   Evaluation metrics for predictive monitoring systems with highly imbalanced datasets,

   supervised by Marlon Dumas and Pavlo Tertychnyi.

2. Publication of the thesis is not allowed.

3. I am aware of the fact that the author retains the rights specified in p. 1.

4. I certify that granting the non-exclusive licence does not infringe other persons' intellectual property rights or rights arising from the personal data protection legislation.

Mariia Godgildieva
13/05/2020