INSTITUTE OF COMPUTER SCIENCE

Academy of Sciences of the Czech Republic

Martin Holeňa

Pod Vodárenskou věží 2, 182 07 Praha 8, phone: +420 266052921, fax: +420 286585789, e-mail: martin@cs.cas.cz
web: www.cs.cas.cz/~martin

Review of the PhD thesis

Mathematical Modeling of Anomalies

in Large-Scale Vector and Structural Data

by Martin Flusser

The reviewed dissertation studies and experimentally validates neural network-based surrogate modelling for anomaly detection. Surrogate modelling is very important in situations where available computational resources do not allow to use a traditional accurate *k*-nearest-neighbors classifier.

The thesis is, apart from the introduction and conclusion, structured into four chapters. Chapter 1 recalls anomaly detection and brings a survey of relevant prior art methods. Chapter 2 presents the basic variant of the author's methodological contribution – surrogate neural network for anomaly detection trained on an auxiliary data set. It is evaluated on the one hand on public data sets, on the other hand on industrial cyber-security data. The chapter pays attention both to detection accuracy and to inference speed, and in addition also to robustness with respect to properties of auxiliary data.

In the remaining chapters, the author's method is elaborated for two specific more complicated settings that are important in cyber-security. In particular in Chapter 3, the setting of online learning from a continuous stream of large volumes of incoming data is considered, whereas in Chapter 4, a multiple instance learning setting. In connection with the respective settings, those chapters pay particular attention to such aspects of the proposed methodology as specific design of the updating procedure, evaluation metrics or combining different strategies.

There are two important reasons why this dissertation deserves to be highly appreciated. First, using the paradigm of surrogate modelling allows to combine the accuracy of distance-based methods with the scalability of neural networks. For the application area of anomaly detection, this approach is a novel solution to the traditional problem of tradeoff between accuracy and computational cost. The other reason is that although the dissertation is a contribution to basic research, it also can have a great practical impact in the application domain of cyber-security.

As far as the methodology of the reported research is concerned, I have only a slight problem with the way how statistical significance of the found differences between compared methods has been assessed. The detection of significant differences by means of confidence intervals is a very crude method, which misses some significantly different pairs of methods that can be detected by specific tests of equivalent performance. And presented results of Wilcoxon test lack familywise correction for multiple hypotheses testing. In a machine-learning dissertation, it is not at all surprising to encounter those methodological problems. Although they have been for many decades known in the area of statistical modelling, the machine-learning community became aware of them less than 20 years ago and till now, they are encountered in some papers even at the most prestigious machine-learning conferences.

From the presentation point of view, the dissertation is well structured and well written although I would sometimes prefer a more detailed explanation of the presented methodology and results. For example, an exact definition of at least some of the aggregation functions considered in Subsection 4.4.7, or an explanation of how the points in Figure 4.11. have been obtained.

After reading the thesis, I had a number of questions to it. I really appreciate that the author was so kind and visited me to answer in detail all of them.

Despite the above mentioned slight methodological and presentation flaws, I find the submitted thesis very good and substantially contributing to the area of anomaly detection. It clearly shows that Martin Flusser can perform creative and systematic research, herewith fulfilling the requirements for a PhD degree.

Prague, August 18, 2023.

Martin Holeňa