UNIVERSITY of GUELPH

School of Computer Science

PhD Defence

Monday, March 26, 2018 at 10:00 AM in MacKinnon Building Room 236

A Neuro-fuzzy Classifier for Datasets with Skewed Feature Values

Jamileh Yousefi

Chair: Dr. Joseph Sawada
Advisory Committee Member: Dr. Charlie Obimbo
Advisory Committee Member: Dr. Rozita Dara
Non-Advisory Committee Member: Dr. Fangju Wang
External Examiner: Dr. Ebrahim Bagheri (Ryerson University)

ABSTRACT:

Most machine learning algorithms perform poorly on skewed datasets. Data distributions in machine learning, when they are discussed at all, are generally expected to have a symmetric distribution, if not actually be normally distributed. The impact of skewed data distribution on the performance of machine learning algorithms has not been given much attention.

Skewed feature values are commonly observed in biological and medical datasets. This poses a challenge for the classification of medical data. Neurofuzzy systems are common machine learning approaches in the medical domain because of their ability to learn fuzzy rules from training data and represent the rules in an understandable way. Therefore, addressing skewness in neurofuzzy systems is a topic of interest because of their applicability in the medical domain.

In this thesis, the Nefclass neurofuzzy classifier is extended to provide improved classification accuracy over the original Nefclass classifier when trained on skewed data. In order to improve accuracy, we used two methods. Firstly, we used two alternative discretization methods. Secondly, we devised several asymmetric linguistic hedges.

The accuracy-transparency trade-off is also one of the most notable challenges when applying machine learning tools in the medical domain. Therefore, the second problem addressed is improving the transparency of Nefclass without significant accuracy deterioration.
We have devised a statistical rule pruning algorithm which uses adjusted residuals to reduce the number of rules, thus improving transparency. Moreover, a hybrid approach combining the above approaches is proposed.

The algorithms have been evaluated on synthetic F-Distributed and Circular-Uniform-Distributed datasets. Additionally, they have been assessed using real-world electromyography and Wisconsin Diagnostic Breast Cancer datasets, which are known to have highly skewed feature values. We evaluated the accuracy of the classifiers using misclassification percentages, and the transparency of the rule-based classifiers using the number of rules. Both independently and in combination, our three approaches provide a considerable improvement in classification accuracy and transparency on skewed data.

This research can contribute to an improvement in decision-making in healthcare or any other area where a significant fraction of the domain data has highly skewed distributions of feature values. In particular, our strategy has led to greater diagnostic accuracy to distinguish neuromuscular diseases using electromyography data which is known to have highly skewed distributions of feature values. This methodology is not limited to Nefclass and neurofuzzy systems because our approaches are not directly tied to the structure of Nefclass. Hence, we expect that our techniques can be applied to any application in which fuzzy logic is used. Furthermore, our rule pruning approach has the potential to be used in other fuzzy and non-fuzzy classifiers.