博士学位論文審査要旨

２０２１年1月２３日

論文題目： Improving Classification Accuracy for Machine Learning
(機械学習における分類精度の向上)

学位申請者： 鄭　弯弯

審査委員：
主　査： 文化情報学研究科　教授　金　　明哲
副　査： 文化情報学研究科　特別客員教授　吉野　諒三
副　査： 文化情報学研究科　教授　沈　　力
副　査： 文化情報学研究科　教授　下嶋　篤
副　査： 文化情報学研究科　教授　宿久　洋

要　　旨：

本論文は、機械学習における高次元小サンプルデータ、クラスの不均衡性とノイズ・異常値の三つの課題について、特徴の選択、分類器の選択とノイズの検出の三つの側面から分類精度の向上を目的とした研究をまとめたものである。その中、特徴の選択とノイズの検出については、新しい方法を提案する内容であり、分類器の選択は実証研究になっている。論文は、５章より構成されている。第１章では、機械学習の現状などを述べたうえ、本研究で扱った三つの課題を提示した。第２章では、小サンプルデータの場合、特徴選択が不安定であり、かつ正確さが欠けている問題点に対して、学習データの代表性と多様性に注目し改善方法を提案した。第３章では、優れていると評価されている計７カテゴリの代表的な１２個の分類器についてチューニングを行い、各分類器の最大のパフォーマンスを発揮させた上で、クラスの不均衡性と学習データサイズの二つのデータ性質が分類器精度に与える影響について検証を行った。第４章では、ノイズが分類器の学習を妨げる問題点に対して、多因子ベース(multi-factor-based)の機械学習に基づいたクラスノイズの高速検出方法を提案した。第５章では、第２章、第３章と第４章の分析の主な結果をまとめ、提案手法の優位性と問題点を示したうえで今後の課題と展望を述べた。

本論文で提案した２つの方法は、実用性を視野に入れ、計算コストと精度の両方面から有効性を示しただけではなく、そのアイディアは今後の研究の発展に寄与するものである。よって本論文は、博士（文化情報学）（同志社大学）の学位論文として十分な価値を有するものと認められる。

総合試験結果の要旨

２０２１年1月２３日

論文題目：　Improving Classification Accuracy for Machine Learning
　　　　　　(機械学習における分類精度の向上)

学位申請者:　鄭　穹穹

審査委員：
主　査：　文化情報学研究科　教授　金　　明哲
副　査：　文化情報学研究科　特別客員教授　吉野　諒三
副　査：　文化情報学研究科　教授　沈　　力
副　査：　文化情報学研究科　教授　下嶋　篤
副　査：　文化情報学研究科　教授　宿久　洋

要　　旨：

学位申請者は２０１８年度４月より本学大学院文化情報学研究科博士後期課程に在学しており、国内会議および国際会議での研究発表を通じて研究活動を積極的に行い、それらの成果を国際論文誌に３本、国際会議 Proceedings に５本の論文として公刊している。また、英語の語学試験にも合格していることから語学（英語）について十分な能力を有していると認定されている。

申請者の学位申請に関し、２０２１年1月２２日金曜日１７:３０から約1時間１５分の公聴会と１０分の審査会において、種々の質疑応答を行った。申請者は研究内容及び関連する質問に対し的確に対応したことで、委員会は申請者が博士（文化情報学）（同志社大学）の学位を授与するに十分な学力を有することを確認した。

よって、総合試験の結果は合格であると認める。

博士学位論文要旨

論文題目：　Improving Classification Accuracy for Machine Learning
（機械学習における分類精度の向上）

氏　　名：　鄭　穹穹

要　　旨：

Machine learning refers to a set of methods that can automatically detect patterns in data, and then use the uncovered patterns to predict future data or other outcomes of interest, or to perform other kinds of decision making under uncertainty. Especially in the era of big data, the information we are drowning in has gained prominence so that it becomes impractical for scientists to handle it humanly. Machine learning enables analysis of massive quantities of data, which can be numbers, words, images, voice and clicks, what have you. That way, it has become an important aspect of modern business and research, and powers many of the services we use today, such as recommendation systems like those on Amazon, Netflix and Watson; search engines like Google, Baidu and Bing; social-media like Twitter, TikTok and Facebook; voice assistants like Siri, Cortana and Nina. This list goes on.

Machine learning is not a new science (since 1949 when D. Hebb created a model of brain cell interaction) but has gained fresh momentum in the current era, and still faces numerous problems. Among them, high-dimension, low-sample-size data, class imbalance and noise/outlier are highly emphasized. This study tried to improve the classification accuracy of machine learning from feature selection, classifier selection and noise detection regarding to the three challenges, respectively.

This thesis is organized under five chapters. Chapter 1 gives a brief explanation of what machine learning is and why it matters. Chapter 2 makes a proposal to improve the performance of feature selection methods with low-sample-size data. Chapter 3 studies the effects of class imbalance and training data size on classifier learning empirically. Chapter 4 proposes a fast noise detector referring to the problems of noise detection algorithms, which are over-cleansing, large computational complexity and long response time. Chapter 5 draws a summary and the closing.

In Chapter 1, first and foremost, the configuration of machine learning was explained, which are supervised learning, unsupervised learning, semi-supervised learning and reinforcement learning. Then, three universal challenges (high-dimension, low-sample-size data; class imbalance; noise/outlier) were stated. In addition, the reasons and the problem tasks were emphasized simultaneously. Finally, the purpose of this study was clarified referring to these three challenges in machine learning.

Feature selection has been a vital topic in the research of text classification, data mining, pattern recognition, and machine intelligence. Given that learning algorithms may be negatively affected by the presence of irrelevant and redundant features, many studies argued that a subset of features may produce better predictive models. Feature selection reduces the dimensionalities of data sets, which helps better understand data, improves the performance

of machine learning techniques, and minimizes the requirement on computation and storage. In Chapter 2, the stability of feature selection was discussed, which recently attracted much attention, especially for the low-sample-size data. According to feature selection methods, sufficient samples are usually required to select a reliable feature subset, especially considering the presence of outliers. Because for low-sample-size data, few outliers can significantly convert the set of selected features into a new set of potential noisy features that may not fully reflect or capture class-specific differences. Furthermore, though conventional feature selection adopts random sampling to improve the performance, low-sample-size data are typically too small to be processed using this method. A lot of previous research tried to increase the sample size to obtain a more stable feature subset, however, this method would increase the computational complexity. This study proposed a method (Feature selection based on data quality and variable training sample, QVT) to improve the stability of feature selection in the extreme case of very small-size data. An experiment was performed using 20 benchmark datasets, three feature selection methods, one synthetic data generation method and three classifiers to verify the feasibility of this method. The results clarified that (1) the feature selection methods fit a classification problem with less than 33 training samples; (2) a smaller number of training samples led to a more significant difference between QVT and the baselines, and QVT was verified as the better one. The classification accuracy was improved up to 13%; (3) QVT fits different feature selection methods, and it can significantly improve the predictive performance of different classifiers; (4) QVT shows to be more robust to handle feature selection for low-sample-size data than the synthetic generation method ANS.

Not only features, In Chapter 3, the impact of two other data characteristics (class imbalance and training data size) on the performance of classifiers were studied. Data characteristics are known to have an intrinsic relationship with classifier performance and choosing appropriate classifiers for a given dataset is very important in practice. Although numerous studies have been conducted on the topic of relationship between data characteristics and classifier performance, there are over twenty most common items of data characteristics and over 100 classifiers have been proposed. Researches perform studies in each field and the reference value of their results is limited. In this study, an empirical study was performed on twelve classifiers arising from seven categories, which are frequently employed and have been identified to be efficient. Furthermore, comprehensive hyperparameter tuning was done for every data to maximize the performance of each classifier. The predictive accuracy and the rank of classifiers were studied and the results indicated that (1) naïve Bayes, logistic regression and logit leaf model are less susceptible to class imbalance while they have relatively poor predictive performance; (2) ensemble classifiers AdaBoost, XGBoost, RF and parRF have quite poorer stability in terms of class imbalance while they achieved superior predictive accuracies; (3) no one superior classifier shows to be robust to the change of training data size; (4) a comprehensive hyperparameter tuning may be able to eliminate the effect of class imbalance.

In Chapter 4, a fast class noise detector with multi-factor-based learning (FMF) was proposed. Noise detection is a preprocessing technique that can be employed in any given dataset to identify potentially noisy instances, which not only offer misleading information and then hinder the learning process of classifiers, but also can significantly alleviate the stability of feature selection. In practice, uncertain or contaminated training sets are commonly conducted. Besides, some studies estimated that even in controlled environments at least 5% of errors exist in a dataset. Thus, detecting class noise prior to the analysis of polluted data appears to be necessary. Noise detection algorithms commonly face the problems of over-cleansing, large computational complexity and long response time. Preserving the original data structure is uttermost important for any classifier. Obviously, over-cleansing will adversely affect the quality of data. Besides, the high time complexity remains one of the main defects for most noise detectors, especially those exhibiting an ensemble structure. Moreover, with numerous studies reported that ensemble-based techniques outperform other techniques in the accuracy of noisy instances identification, these problems are scaling up. This study proposed a fast class noise detector with fundamentally different structure that makes a multi-factor-based learning rather than the normal simple learning (e.g., density-based learning) or ensemble-based learning with classifiers. According to the results, (1) removing instances identified by FMF for training achieved the highest overall classification accuracy compared with the classifiers trained on the entire training data as well as with noise removed by the other methods. The classification accuracy was improved up to 18%; (2) FMF efficiently sped up the computation. Compared with the second fast dynamicCF and the slowest INFFC, its speed was about 20 times and 333 times faster, respectively; (3) FMF reached the lowest reduction rate. Compared with All *k*-NN which achieved the maximum reduction rate, the value of FMF was about its 1/3. However, the low reduction rate also demonstrates the low possibility of over-cleansing.

In Chapter 5, a summary was drawn. In detail, the studies of Chapter 2 (feature selection), Chapter 3 (classifier selection) and Chapter 4 (noise detection) were described from background, problems and highlights, respectively.