氏　　名	HOANG, Anh
学位の種類	博士（知識科学）
学位記番号	博知第 294 号
学位授与年月日	令和 3 年 9 月 24 日
論文題目	Outlier Detection and Class Imbalance Based on Mass Estimation Integrated with Evidential Reasoning
論文審査委員	主査　HUYNH Van Nam　北陸先端科学技術大学院大学　教授
	橋本　敬　同　教授
	DAM Hieu Chi　同　教授
	由井薗　隆也　同　教授
	LE Hoai Bac　University of Science, VNU- HCM　教授

論文の内容の要旨

Outlier detection and class imbalance modeling process play significant roles to enable effective and efficient algorithms for statistic analysis, data mining, machine learning, and knowledge discovery frameworks working on imbalanced datasets. Although there has been vast literature on imbalanced datasets, the shortcomings of distance-based functions in response to a varied density of data points have not been solving yet.

The primary aim of this dissertation was to exploit a new alternative approach for local outlier detection tasks by fundamentally changing the way to measure the outlier degree of each data point. To achieve this goal, we developed a mass-based approach to measure the dissimilarity between data points. Then, we introduced a new outlier scoring method by employing mass-based dissimilarity and probability modeling to detect the local outliers in a given dataset. The experimental study tested on artificial datasets and real application datasets show that our proposed MLOS approach is competitive with the state-of-the-art approaches.

In the same manner, to exploit the mass-based measurement for learning from the imbalanced datasets, we introduce the other two new methods for the class imbalance task. The first model is a simple application of weighted sum. The second model is an integration of the mass estimation and the Dempster-Shafer theory of evidence. These proposed models were assessed by using significant evaluation metrics such as F1 score, Brier score, ROCAUC, and PR-AUC score testing on a wide range of benchmark datasets. In addition, all experimental results were validated using the non-parametric statistical Wilcoxon signed ranks test.

This dissertation was the first study, regarding to our knowledge, to investigate the local outlier detection problem using mass-based dissimilarity measurement; the key finding was that the proposed MLOS approach presents an alternative way to score the outlierness of each data point in a given dataset. Secondly, the simulation results showed that our proposed new models for the class imbalance task outperformed the other 11 competitive methods. The experiments were conducted on a wide varying application domains, a varied imbalance ratio, and the number of instances.

Keywords: Imbalanced data, outlier detection, outlier modeling, mass-based dissimilarity, weighted sum, Dempster-Shafer theory.

論文審査の結果の要旨

The problems of outlier detection and class imbalance classification play significant roles in many fields of statistical analysis, data mining and machine learning. These problems have been addressed by both unsupervised and supervised approaches. The primary objective of this research is to exploit a new alternative approach for outlier detection and class imbalance classification tasks by using mass-based dissimilarity measures instead of distance-based measures in combination with evidential reasoning in terms of the Dempster-Shafer theory of evidence. The main results of this research are summarized as follows.

As for the problem of outlier detection, this dissertation first proposes a mass-based approach to measure the dissimilarity between data points and then develops a new outlier scoring method, namely mass-based local outlier score (MLOS) method, by employing mass-based dissimilarity and probability modeling to detect the local outliers in the dataset. The experimental study tested on artificial datasets and real-world datasets show that the proposed MLOS method is competitive with various state-of-the-art existing methods.

As for the problem of class imbalance classification, two new models for classifying imbalanced datasets are developed in this dissertation. The first model is called mass-based similarity weighted k-neighbor model that is based on the concept of k-neighbors defined using mass-based similarity measure. The second model is an integration of the mass estimation and the Dempster-Shafer theory of evidence. These proposed models were experimentally tested on a wide range of benchmark datasets using significant evaluation metrics such as F1 score, Brier score, ROCAUC, and PR-AUC score. In addition, all experimental results were also validated using the non-parametric statistical Wilcoxon signed ranks test.

This dissertation has made good contributions to methodological and experimental developments within the areas of outlier detection and class imbalance classification. The research work presented in this dissertation has resulted in two journal papers and two refereed conference papers.

In summary, Mr. HOANG Anh has completed all the requirements in the doctoral program of the School of Knowledge Science, JAIST and finished the examination on August 2, 2021, all committee members approved awarding him a doctoral degree in Knowledge Science.