Ministry of Education of the Republic of Belarus
Educational Institution
Belarusian State University
of Informatics and Radioelectronics

UDC 004.94

Guo Qiang

IMBALANCED DATA CLASSIFICATION ALGORITHM

ABSTRACT
for a master's degree

Speciality 1-40 80 02 System analysis, information control and processing
(by industries)

Academic Supervisor
German Oleg Vitoldovich
Candidate of Technical Sciences,
Associate Professor of the
Department of ITAS, BSUIR

Minsk 2024

INTRODUCTION

With the rapid development and popularization of computer technology and the Internet in recent years, massive amounts of data are generated in various fields every day. These data come from various sources and are complex and disorderly. On the one hand, they contain the information that people hope to obtain, and on the other hand, they are full of useless information that even interferes with normal judgment. In order to obtain hidden useful information, data mining technology has become a focus of scientific researches. It has been developed rapidly in recent years and many new methods have emerged. Classification is an indispensable step in data mining. Data classification problem is an important research direction in the field of machine learning. Its main purpose is to use previously obtained data with specific category labels to train a mathematical model so that it can be used after debugging and training when encountering similar unknown category data later. Good mathematical models assist classification and decision-making based on the principle of similarity between labeled data. Mathematical models that have been widely used include support vector machines, decision trees, neural networks and other models. Through the research and application of data classification methods, people have been able to automate some life problems, such as automatic recognition of license plate numbers, banknote recognition, etc.

Although the current research on data classification problems has become increasingly mature, it still encounters major challenges in practical applications. Among them, the imbalance between the obtained data samples is one of the thorny problems. Most traditional machine learning algorithms are proposed on the premise that each category of data basically meets the balance. However, when it comes to specific research problems such as fault diagnosis, medical diagnosis, text classification, credit card fraud detection, etc., the data sets we obtain often have varying degrees of imbalance. The imbalance will make it easier for the classifier to classify test samples into the majority class due to excessive bias in traditional machine learning algorithms. Therefore, it is particularly necessary to use some data processing methods or improve traditional classification algorithms to adapt to data imbalance scenarios.

The problems faced by imbalanced data classification are divided into the following categories:

– Inappropriate evaluation criteria. In the field of data mining, appropriate evaluation criteria are needed. There are many classification algorithms that build classification models under the guidance of evaluation criteria. For example, decision trees use the information gain criterion as the construction criterion, and

ultimately use classification accuracy to evaluate the overall classification performance.

– Absolute scarcity of samples. The absolute scarcity of samples means that the number of minority class samples themselves is very small, and the information they contain is naturally very scarce, making it difficult to discover the distribution pattern of minority class samples, seriously affecting the final classification performance.

– The relative scarcity of samples. The relative scarcity of samples means that the absolute number of minority class samples is not small, but compared with the number of majority class samples, it is very small. In this case, if a single condition is used to distinguish different classes, even if the minority class samples are indeed satisfied, due to their relative scarcity, they will be submerged in the majority class samples, or it can be said that the decision boundary between the majority class samples and the minority class samples will therefore become unclear.

– Data fragmentation. There are many classification algorithms that will continuously decompose the original problem into several small sub-problems to be solved respectively. Each small sub-problem is only based on a certain small space segment divided from the original sample space, which is the so-called data fragmentation. Each small sub-problem can only be processed independently by using data fragments to find patterns. When facing minority class samples, such problems will be more prominent, because the originally sparse minority class samples have been divided into many data fragments. There is no pattern at all found in it. Therefore, classification algorithms using this “divide and conquer” strategy will face more serious difficulties when faced with imbalanced data classification.

– Improper inductive bias. Usually, classification algorithms or other data mining algorithms hope to obtain better generalization capabilities and reduce the occurrence of overfitting, so another evidence-based bias is needed when generalizing a certain sample. This principle is important for the generalization performance of the algorithm. The number of minority class samples is small, and it is difficult to often find multiple biases with evidence, which makes the patterns in minority class samples easily overlooked. Therefore, this inductive bias method is not suitable for minority class classification.

– Influence of noise samples. Noise samples will inevitably appear in classification algorithms, and will have a negative impact on almost all classification algorithms. When faced with imbalanced data classification, it is conceivable that due to the scarcity of minority class samples, the impact of noise samples will sometimes be very significant, because the information and rules contained in minority class samples are very weak. A few noise samples will cause considerable deviations in the distribution of this type of samples.

The core requirement of imbalanced data classification is how to effectively improve the classification effect of minority class samples. Most traditional classification algorithms are designed based on the assumption of balanced data, and it is easy to sacrifice the recognition rate of minority class samples for the overall recognition rate. However, in practical applications, it is difficult to ensure a balanced distribution of samples of different classes, and when data imbalance occurs, it tends to focus on the information contained in minority class samples. When most traditional classification algorithms deal with imbalanced data classification, they will favor majority class samples to improve the classification accuracy of the overall sample. The purpose of the imbalanced data classification method is to improve the minority class samples that are more important in practical applications, improve the classification accuracy and improve the overall classification performance.

In this context, this thesis briefly introduces several currently popular solutions, and through simulation experiments to compare the advantages and disadvantages of these solutions; on this basis, an algorithm to solve the data imbalance problem is proposed and verified through simulation experiments.

GENERAL DESCRIPTION OF THE WORK

Relevance of the subject

This research is relevant because it can improve classifier performance, enhance recognition of minority classes, and aid decision-making. This is crucial for the application of artificial intelligence in various fields. The dissertation research was carried out within the framework of the direction “Digital information and communication and interdisciplinary technologies, production based on them” on the realization and implementation of hardware and software solutions using artificial intelligence, big data bases for the Internet of things, industrial Internet, cloud technologies, intelligent electronic terminals in accordance with the State Program for Innovative Development of the Republic of Belarus for 2021–2025, approved by Decree of the President of the Republic of Belarus dated September 15, 2021 No. 348.

The aim and tasks of the work

The aim of the work is to improve the minority class samples that are more important in practical applications, improve the classification accuracy and the overall classification performance.

To achieve this aim, the following tasks were solved in the dissertation:

1. Several common imbalanced data classification methods have been studied, including random upsampling, random downsampling, and neighborhood weighted K-nearest neighbor algorithms. Experimental simulations were conducted for these four algorithms, and the results were analyzed.

2. Proposed weighted clustering based on PDF, experimental simulation was conducted on the algorithm, and the results were analyzed.

3. Proposed PDF-based KNN algorithm for imbalanced data, experimental simulations were conducted on the algorithm, and this algorithm greatly improving the classification accuracy of minority class samples, shows stable performance on different data sets, and achieves very ideal results.

Personal contribution of the author

The evaluation criteria for imbalanced data and imbalanced data classification are introduced, and the AUC-ROC standard is cited as the evaluation criterion for classification performance. Several common imbalanced data classification methods have been studied, including random upsampling, random downsampling, and neighborhood weighted K-nearest neighbor algorithms. Experimental simulations were conducted on these three algorithms, and the results were analyzed. This thesis proposes an imbalanced PDF data processing method based on Gaussian distribution, conducts simulation experiments, analyzes the results, and proposes a classification method based on PDF weighted clustering, which can effectively improve the performance of the classifier. However, the results are less than ideal due to sensitivity to noise and outliers. Therefore, an imbalanced data classification method based on PDF-based KNN algorithm is proposed. This algorithm greatly improves the classification accuracy of minority class samples, shows stable performance on different data sets, and achieves very ideal results.

Task setting and discussion of the results were carried out together with the associate professor of the ITAS department, German O.V.

Testing and implementation of results

The main provisions and results of the dissertation work were reported and discussed at: On approval of topics of master’s theses and scientific supervisors of master’s students for the 2022 intake.

The results of the thesis were used in educational process in ITAS Educational institution Belarusian State University of Informatics and Radioelectronics in the course Systems of Analytical Programming (lecturer dr. German O.V.)

Author's publications

According to the results of the research presented in the dissertation, 2 author's works were published, including: 2 articles and abstracts in conference proceedings.

Structure and size of the work

The dissertation work consists of introduction, general description of the work, three chapters with conclusions for each chapter, conclusion, bibliography, appendix.

The total amount of the thesis is 76 pages, of which 53 pages of text, 17 figures on 7 pages, 3 tables on 1 page, a list of used bibliographic sources (39 titles on 2 pages), a list of the author's publications on the subject of the thesis (2 titles on 1 pages), 1 appendix on 7 pages, graphic material on 5 pages.

SUMMARY OF THE WORK

The introduction discusses the current research status of imbalanced data classification algorithms, analyzes the problems faced by the algorithm and the necessity of continuing to study and learn imbalanced classification algorithms. It also puts forward the main work objectives of the paper and briefly introduces the chapter structure of the paper.

The general description of work shows the connection between the work and the priority areas of scientific research, the aim and tasks of the research, the personal contribution of the applicant for a scientific degree, the approbation of the dissertation results.

In the first chapter introduces several common imbalanced classification evaluation criteria, including F-measure, G-mean and AUC-ROC. And AUC-ROC was selected as the algorithm evaluation standard in the simulation experiment of this paper. We also studied several imbalanced data classification methods, including resampling methods and neighborhood weighted K nearest neighbor algorithms, and conducted experimental simulations of these methods. Analyzing the experimental results found that for the classification of imbalanced data, there are some problems with traditional classification algorithms. Serious flaws, so methods for handling imbalanced data are open.

In the second chapter, the PDF-based imbalanced data classification method is introduced and experiments are conducted. The results obtained were not ideal, and a new method was proposed, the PDF-based KNN algorithm. This algorithm combines the Gaussian mixture model and the KNN algorithm, uses GMM to

estimate the sample distribution of the majority class and the minority class respectively, and then uses the minority class distribution differences between and majority classes to generate new samples. This method reduces the focus on the majority class of the data set and solves the problem of class imbalance. The simulation experiment results show that the PDF-based KNN algorithm effectively improves the classification performance of imbalanced data and is an effective method for classifying imbalanced data.

In the third chapter introduces the programming tools used in all simulation experiments in the paper, and also introduces the main codes and explanations of the experiments, as well as the user manual. In order to prevent the algorithm proposed in this article from being concretized in a single data set and to verify the generalization ability of the model, this chapter refers to other public data sets on sklearn and compares their final results. The simulation experiment results show that the imbalanced data classification method based on the PDF-based KNN algorithm effectively improves the classification accuracy and overall classification performance of minority class samples, and shows strong stability on different data sets. The PDF-based KNN algorithm is a method suitable for imbalanced data classification.

CONCLUSION

In today's society, people are receiving more and more information, and it is becoming more and more complex. There is an urgent need to quickly and accurately find useful information in massive data. Therefore, data mining technology is attracting increasing attention. Classification is one of the important knowledge acquisition methods in data mining and machine learning. Classic classification algorithms are usually proposed based on the assumption that the data set is balanced. In real life, many data sets are unbalanced, and smaller amounts of data are often more important. Therefore, traditional classification algorithms that take overall classification accuracy as the learning goal are not suitable for classifying imbalanced data.

Due to the widespread existence of imbalanced data classification in reality and the theoretical challenge of the problem itself, it has attracted many scholars to invest in related research. So far, many methods and theoretical analyzes have been proposed, but due to the complex sources and different characteristics of imbalanced data sets, it is difficult to have a single algorithm that can show excellent performance in all situations, and in practical applications There are also a large

number of imbalanced data classification problems that need to be solved urgently. Therefore, further research on imbalanced data classification methods still has practical and specific significance. In the application of traditional classification algorithms, it is usually assumed that the data sample set is balanced. However, due to the data imbalance problem that often exists in actual application scenarios, ideal classification results are usually not achieved. In order to solve the problem of inaccurate minority class sampling ratio in imbalanced data classification and insufficient attention to minority class samples in the algorithm, the difference between the majority class and the minority class is used to generate the same number of minority classes as the majority class. This thesis proposes a PDF-based KNN algorithm. The main results of this thesis are summarized as follows:

1 There were introduced the evaluation criteria of imbalanced data and imbalanced data classification, and cited the AUC-ROC criterion as the evaluation criterion of classification quality. Several common imbalanced data classification methods were studied, including random upsampling, random downsampling, and neighborhood weight K nearest neighbor algorithms. Experimental simulations were developed for these three algorithms, and their results were analyzed.

2 There was proposed the method of processing unbalanced PDF data based on Gaussian distribution, formulated simulation experiments with analyses of the results, and proposed a classification method based on weighted clustering utilizing PDF, which could effectively improve the performance of the classifier. However, due to sensitivity to noise and outliers, the results were less than ideal. Therefore, there was proposed an imbalanced data classification method utilizing the PDF-based KNN algorithm. Use GMM to obtain the sample distributions of the majority class and the minority class respectively, then calculate the difference between the distributions of the minority class and the majority class, and use the difference to generate new samples with the same number as the majority class, which belong to the minority class. This algorithm greatly improves the classification accuracy of minority class samples, shows stable performance on different data sets, and achieves very ideal results.

3 There were introduced the programming tools and formulated the user manual. In order to prevent the algorithm from accidental occurrence, 5 public data sets with different balances were quoted to realize simulation experiments and analyze the experimental results. Simulation experiments showed that the algorithm proposed in this thesis effectively improved the accuracy of imbalanced data classification and showed stable performance on different data sets.

The problem of imbalanced data classification arises in specific practice. Given thesis studies and improves many traditional classification algorithms. However, due to limitations of hardware conditions and author’s own abilities, the

results of this thesis still need to be improved in the forgoing research, specifically divided into the following aspects:

1 The method proposed in this thesis only uses Euclidean distance, which is not necessarily the most appropriate choice for many imbalanced data sets. Therefore, in the next step of research work, single-class classification ideas and distance metric learning ideas can be introduced to improve the classification performance of imbalanced data classification.

2 Potential risk of overfitting: By focusing only on the synthesis of minority class samples, the risk of overfitting may be introduced. The generated synthetic samples may be too adapted to the characteristics of minority class samples and thus perform poorly on real test data.

LIST OF PUBLICATIONS OF THE APPLICANT

1–A. Guo Qiang. Imbalanced data classification algorithm / Qiang Guo, O. V. Geman // Information Technologies and Systems 2022 (ITS 2022): Proc. of the International Conference, 23 November 2022 / BSUIR, Minsk – 2022. – P. 151–152.

2–A. Guo Qiang. Smote algorithm in imbalanced data / Guo Qiang // 59th conference of postgraduate students, undergraduates and students of the educational institution "Belarusian State University of Informatics and Radioelectronics", 17-21April 2023 / BSUIR, Minsk – 2023. – P. 59–60.