IMBALANCED DATA CLASSIFICATION ALGORITHM

Guo Qiang, German O.V.
Department of Information Technology in Automated Systems,
Belarusian State University of Informatics and Radiolectronics
Minsk, Republic of Belarus
E-mail: 254951872@qq.com, ovgerman@tut.by

This article mainly analyzes the characteristics and challenges of imbalanced data, analyzes the research status of imbalanced data integration classification algorithms and combines the actual situation, based on ensemble learning. Research and improvement of classification algorithms.

INTRODUCTION

In the modern society with highly developed big data and the Internet, based on the diversity and complexity of data information, imbalanced data widely exists. Therefore, the problem of imbalanced data classification will be an insurmountable difficulty in the data field, and it has very important practical value for its research and learning. Classification is one of the important knowledge acquisition methods in machine learning and data mining, and existing classification algorithms usually assume that the data set used for training is balanced. But real-world classification problems are often not balanced datasets, and even have serious imbalances. When encountering class data imbalance, the traditional classification algorithm with the overall classification accuracy as the learning goal will pay too much attention to the majority class, and the classification performance of the minority class samples will be degraded, such as oil leak detection, medical diagnosis, software fault detection, etc. [1]

Imbalanced data classification focuses on the performance of learning algorithms when the class data is unbalanced or under-represented, and the study of this problem has become one of the hot topics in the field of machine learning. According to the existing research results, solving the classification problem of imbalanced data can be handled by rebalancing the data through sampling techniques.

However, the ensemble learning algorithm aims to reduce the overall classification error rate, but does not consider the difference in the classification cost of different samples, and often has a poor recognition effect on key minority samples. Therefore, the improvement of the ensemble learning algorithm is also a great challenge.

I. Classification algorithm based on ensemble learning

Slave-balanced data refers to the huge difference in the number of samples in each category of the data set. Taking binary classification as an example, the number of samples in the positive class is much larger than the number of samples in the negative class, as shown in Figure 1.

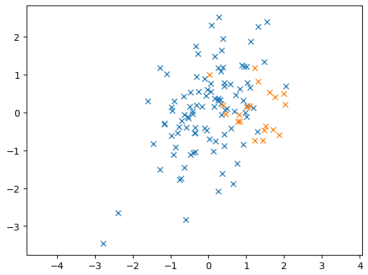

Figure 1 – Imbanlanced data

Imbalanced data are mainly manifested in two types: inter-class imbalance and intra-class imbalance. Inter-class imbalance refers to the unequal distribution of data between classes. In some practical applications, the data presents extreme class-to-class data imbalance, and some imbalance ratios can reach 1000:1 or higher. Intra-class imbalance refers to the imbalance in the number of samples of a class and its subclasses, or the data of a certain class presents multiple small discrete items. Usually, the data imbalance between classes is not the only factor that affects classification learning, and the data imbalance within a class is the key factor that affects the classification effect. Therefore, the problem of imbalanced data classification is mainly due to the complexity of data distribution, such as data overlap, the existence of subclasses of minority classes, and the problem of small separation items. These problems directly affect the learning of the classifier. [2]

The evaluation indicators for imbalanced data are special. First, the minority class concerned by the classification learning task is defined as positive class T, and the majority class is defined as negative class N, so there are four kinds of prediction results: positive and correct (Tp), positive and wrong (Tn), negative and correct (Fn), negative class and false (Fp). The accuracy rate (P) indicates that the positive examples are correctly classified and is sensitive to the distribution of the data. The recall rate (R), also known as the recall rate, indicates the completeness of the correct classification of positive examples and is not sensitive to the distribution of the data. The ideal classification result is that the higher the precision and recall, the better. The F value is the weighted harmonic mean of precision

and recall, and is the most common F1 value when the parameter b = 1. F1 combines the results of precision and recall. When the F1 value is high, the result is ideal. G-means represents the balance between the positive classification accuracy and the negative classification accuracy.

$$P = \frac{Tp}{Tp + Fp} \tag{1}$$

$$R = \frac{Tp}{Tp + Fn} \tag{2}$$

$$F = \frac{\left(b^2 + 1\right) * P * R}{b^2 * R + 1} \tag{3}$$

$$G = \sqrt{\frac{Tp}{Tp + Fn} * \frac{Tn}{Tn + Fp}} \tag{4}$$

As a typical representative of current machine learning, ensemble learning algorithm can improve the overall classification accuracy by group decision-making, and is widely used in the classification of imbalanced data. Ensemble learning is a technique that improves the final learning effect by training and integrating multiple weak classifiers. The multi-sampling technique in ensemble learning can be integrated with the optimal class distribution of imbalanced data and the search for optimal class representative examples to avoid additional learning costs. Ensemble classification learning is a machine learning technology that integrates multiple base classifiers to make decisions together, obtains multiple different base classifiers by invoking a simple classification algorithm, and then combines the base classifiers into one classifier in some way. The accuracy and difference of the base classifier are two important factors that affect the learning effect of the final ensemble classifier. [3]

The ensemble classification algorithm aims at the overall learning accuracy and cannot be directly used for classification learning of imbalanced data, but is processed at the algorithm or data level. Algorithmic processing refers to introducing a cost factor in the training process of the ensemble classification algorithm, and assigning different cost factors according to the different costs of different classes of samples being misclassified to form a cost-sensitive ensemble classification algorithm. Data processing refers to incorporating sampling techniques that rebalance the data in the process of building a base classifier, so that the ensemble algorithm can build a classifier on balanced training data that does not affect learning performance.

In order to solve the insufficiency of the single evaluation index, the ROC curve evaluation index appeared to classify and evaluate the imbalanced data. The ROC curve is a graphical way to show the tradeoff between true and false positive rates of a classification model.The ROC curve is shown in Figure2.

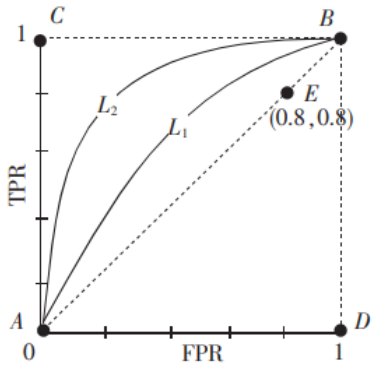

Figure 2 – Roc curve

The area under the ROC curve provides another way to evaluate the average performance of the model, the larger the area, the better the model. An ideal model has an area value of 1. [4]

The classification algorithm based on ensemble learning is very effective in dealing with imbalanced data, but the classification algorithm will never be perfect, there will always be new problems and new requirements, and then the algorithm must be improved and improved, so Further, the imbalanced data classification algorithm has a very broad room for improvement.

II. CONCLUSION

The classification of unbalanced data is widely used in social life, such as credit card fraud detection, spam filtering, and network intrusion detection, which makes the study of this type of data a research hotspot in the field of data mining. This paper reviews the imbalanced data classification algorithm based on ensemble learning. Data distribution adjustment is the most direct method to solve the problem of class imbalance, and the imbalanced data is converted into balanced data to eliminate the impact of class imbalance. Transformation The main methods are resampling, data grouping and single-class learning. Data distribution adjustment changes the distribution of the original data, which may lead to loss of valuable information or overfitting problems.

III. References

1. CHAWLA N. V., JAPKOWICZ N., KOTCZ A. Editorial: special issue on learning from imbalanced data sets [J]. ACM SIGKDD Explorations Newsletter, 2004, 6(1) : 1-6
2. PEARSON R., GONEY G., SHWABE R. J. Imbalanced clustering for microarray time-series [C] //Proc of the 20th International Conference on Machine Learning. 2003.
3. TING K. M. A comparative study of cost-sensitive boosting algorithms [C] //Proc of the 17th International Conference on Machine Learning. 2000: 983-990.
4. DAVIS J. GOADRICH M. The relationship between precision-recall and ROC curves [C] //Proc of the 23rd International Conference on Machine Learning. New York: ACM Press , 2006: 233-240.