A Review on Classification Algorithm for Imbalanced-Class Datasets

Puja Dwivedi
pujadwivedi10@gmail.com
Central College of Engineering and Management
Raipur, Chhattisgarh, India

Dr. P. Udaya Kumar
uday.uday08@gmail.com
Central College of Engineering and Management
Raipur, Chhattisgarh, India

Abstract

The main concentration of traditional classification algorithms addresses imbalanced-class dataset that mostly focuses on the majority classes accuracy, where the minority class's accuracy is usually ignored. In different fields, there is an enormous amount of high-dimensional class-imbalanced data. In this case, traditional classification methods are not appropriate because they are prone to ensure the accuracy of the majority class. To handle imbalanced distribution Re-sampling of imbalanced data is commonly used, as it is independent of the classifier being used. But sometimes they can cause over-fitting. The main focus of work is to gain advantages of both data level and classifier ensemble approach so as to achieve improvement in the classification performance. We present an approach that reduces the imbalance between the classes. When compared with four algorithms, our experimental evaluations have shown that our algorithm can improve the accuracy of a minority class.

Keywords—*Classification,Re-sampling,classifier ensemble,class-imbalanced data.*

I. INTRODUCTION

Classification is one of the important tasks in machine learning and it is used to categorize the data. Many real world classification systems such as fraud detection and credit scoring system face the problem of the Imbalanced dataset where one class consist of a very high number of instances relative to the other class. Hence imbalanced learning is gaining a great deal of attention of researchers in the domain of machine learning.

Many classification algorithms are facing challenges with the emergence of new data types, even though they used to be successfully adapted in different fields. High dimensional and class-imbalanced data are such a typical new data type, which is ubiquitous in various fields, such as biomedicine, cancer diagnosis using DNA microarray data, and image classification. Traditional classification algorithms are based on class balanced hypotheses. When the class label is imbalanced, the standard performance computing method will lead classifier to ensure the accuracy of the majority class by scarifying the minority class. But in most of the cases, the accuracy of minority class is what we should really focus on. For example, in the cancer diagnosis, one of the research focuses was the insurance of the accuracy of predicting a cancer patient. Current researchers still struggle with the complicated situation caused by the real data, although they have paid more attentions to the class-imbalanced problem in recent years. Usually, high-dimensional feature always accompanies with the class-imbalance. Too many variables in the data can result in the "curse of dimensionality". The Feature Selection is a common method that can reduce features and sampling as well as balance the diversity class instance numbers. However, the challenge is that determining the priority between the feature selection and sampling is still difficult since the effects of high-dimension and class-imbalance are mutually infiltrated. Whether or not the priority of preprocessing is associated with the dimension and class-imbalance level of data is still unclear. All these problems have driven researchers to explore in preprocessing.

Such imbalanced class datasets differ from balanced class datasets not only in the skewness of class distributions but also in the increased importance of the minority class. Despite their frequent occurrence and huge impact in day-to-day applications, the imbalance issue is not properly addressed by many standard machine-learning algorithms, since they assume either balanced class distributions or equal misclassification costs. Various approaches have been proposed to deal with imbalanced classes in order to cope with these issues, including over/undersampling, SMOTE (Synthetic Minority Oversampling Technique), cost sensitive, modified kernel-based, and active learning methods. Although these methods do somewhat alleviate the problem, they are often based on heuristics rather than on clear guidance. For instance, in the oversampling technique, the optimal oversampling ratio varies largely across datasets and is usually determined through multiple cross-validation or other heuristics.

Traditional classification algorithms mainly focused on assuring the accuracy of majority class, which lead to lowering the accuracy in predicting the minority class. However what we really need is predicting the minority class correctly. Ensemble learning, that combines the result of several simple classifiers, not only promise the accuracy of the classifier, but also alleviate the overfitting problem. But it also faces the problems as well as the traditional classification algorithms when the imbalanced class datasets are handled. The Ensemble Feature Selection (EFS), a type of ensemble learning approach that is based on feature selection that is promoted by Opitz. EFS differs from traditional feature

selection in a way that EFS does not only aims to find one appropriate set of learning but gets the best ensemble from the ensemble perspective. The performance of the final ensembles is usually strongly affected by the selected feature subsets. The ensemble will be beneficial to the minority class if selected subsets are prone to the minority class. To solve the imbalanced class data classification problem, we propose a classification algorithm based on EFS.

There are essentially three steps in EFS algorithms: 1) Selections of feature subsets. 2) Generating base classifiers. 3) Ensemble prediction results. Fig. 1 represents an illustration showing a framework of EFS.

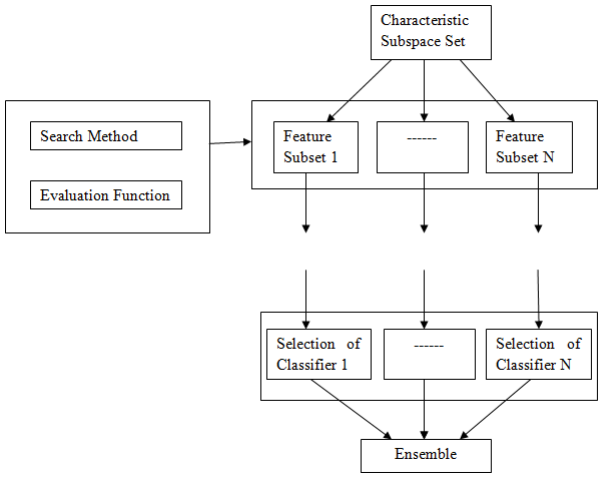

Fig. 1. The framework of ensemble feature selections

II. LITERATURE SURVEY

A. *A Classification Algorithm Based on Ensemble Feature Selections for Imbalanced-Class Dataset*

Opitz proposed EFS which is also called Feature-Based Ensemble Classification. It constructs different classifiers on the distinct datasets according to the selected feature subsets. Since the training dataset derives from different feature subspaces, there will be some diversity factors between classifiers. The diversity factors can aid to construct proper classifier ensembles. Moreover, EFS is different from other traditional feature selection methods that use a single criterion to select a group of optimal or suboptimal features. EFS aims to construct base classifiers with a higher-level accuracy and ensures the diversity of misclassifications, instead of targeting at a certain optimal or suboptimal feature subset. For reaching this goal, it is important to balance diversity and accuracy levels rather than seeking the optimal feature subset when designing the feature selection algorithms, since diversity is usually conflict with accuracy. In this paper algorithm proposed is Imbalanced Ensemble Feature Selection, using HC-based search method.

In this section, we present a few comparative experimental results and display our research findings. We choose three datasets from UCI machine learning repository. They are respectively Steel Plates Faults, Stalog (Landsat Satellite) and Semeion handwritten digit datasets. These three datasets are not imbalanced-class datasets, but they have many classes. We transferred it to a two-class problem and then constructed several imbalanced-class datasets. At last, we choose five datasets in our experiments. They are Steel-Dirtiness, Steel-Pastry, Steel-Bump, Stalog-4 and SemeionA-8. The ratio of majority class and minority class are showed in Table I. We consider both AUC assessment and TPRate to compare four algorithms, including IEFS, CSRF, C4.5, and RF. Table II and Fig. 2 represents the comparisons of the AUC for four algorithms running on different testing datasets. Moreover, Table III and Fig. 3 represents the comparisons of the TPRate for four algorithms running on different testing datasets.

TABLE I. DATASETS

Dataset Name	Minority Class(%)	Majority Class(%)
SteelDirtness	2.83	97.17
SteelPastry	8.14	91.86
SteelBumps	9.36	90.64
Stalog4	9.73	81.27
SemeionA-8	20.7	79.3

TABLE II. COMPARISONS OF AUC FOR IEFS, CSRF, C4.5, AND RF

Dataset Name	C4.5	CSRF	RF	IEFS
SteelDirtness	0.83	0.9	0.85	0.86
SteelPastry	0.78	0.86	0.83	0.9
SteelBumps	0.76	0.9	0.79	0.85
Stalog4	0.79	0.88	0.85	0.86
SemeionA-8	0.81	0.88	0.84	0.85

TABLE III. COMPARISONS OF TPRATE FOR IEFS, CSRF, C4.5, AND RF

Dataset Name	C4.5	CSRF	RF	IEFS
SteelDirtness	0.506	0.764	0.527	0.725
SteelPastry	0.42	0.668	0.291	0.702
SteelBumps	0.354	0.67	0.465	0.64
Stalog4	0.55	0.607	0.536	0.598
SemeionA-8	0.478	0.703	0.357	0.8

IEFS had better performances than C4.5 and RF for both AUC and TPRate evaluations, according to the experimental results. This result had proved that by giving more attentions IEFS could increase the recognition rates of minority classes. When compared with CSRF, IEFS had a slightly lower performance in classification accuracy but it performed better in other aspects. The selected searching method was impacted by these results. A local optimum searching method is Hill climbing searching, which cannot achieve the global optimum. Therefore, the output ensemble classifiers might not be the optimal solutions for all situations.

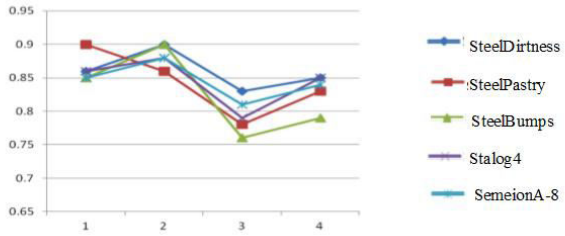

Fig. 2. Comparisons showing AUC curves of IEFS, CSRF, C4.5, and RF

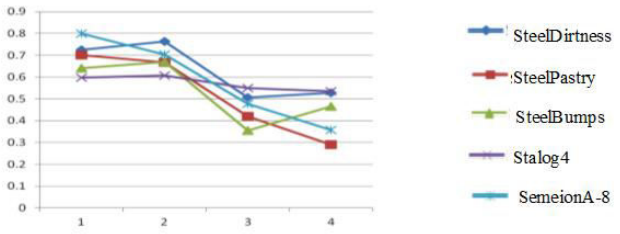

Fig. 3. Comparisons showing TPRate curves of IEFS, CSRF, C4.5, and RF

Conclusion

This paper proposed a new algorithm, Imbalanced Ensemble Feature Selection algorithm, which was designed to solve the low accuracy problem of minority classes when classifications were executed on an imbalanced class dataset. We choose KW variance as the diversity metrics and introduced a punishment-reward factor to dynamically adjust the effect of the feature selections in order to increase the minority classes' accuracy.

B. An Empirical Study on Preprocessing High-dimensional Class-imbalanced Data for Classification

Hua Yin and Keke Gai presented, In the context of software defect prediction, used two data preprocessing steps, feature selection (for selecting the important attributes) and data sampling (for addressing class imbalances), together in the context of the software defect prediction. According to the sequence of these two preprocessing technologies, they considered four possible scenarios: (1) feature selection based on original data, and modeling based on original data; (2) feature selection based on original data , and modeling based on sampled data; (3) feature selection based on sampled data, and modeling based on original data; and (4) feature selection based on sampled data, and modeling based on sampled data.

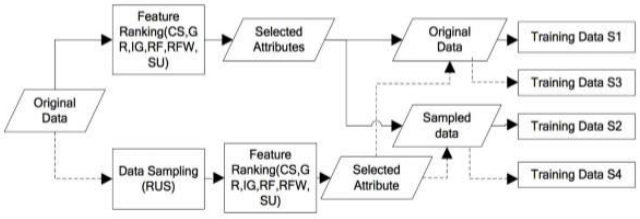

Fig. 4. Four Scenarios of using Feature selection and Data Sampling together

In the context of software defect prediction, used two data preprocessing steps, feature selection (for selecting the important attributes) and data sampling (for addressing class imbalances), together in the context of the software defect prediction. According to the sequence of these two preprocessing technologies, they considered four possible scenarios: (1) feature selection based on original data, and modeling based on original data; (2) feature selection based on original data , and modeling based on sampled data; (3) feature selection based on sampled data, and modeling based on original data; and (4) feature selection based on sampled data, and modeling based on sampled data. Fig.4. shows the four scenarios.

Conclusion

This paper presents four conclusions: (1)feature selection firstly is a little better than sampling firstly. (2) when the dataset is largely imbalanced, undersampling is more useful. (3) when the dataset is less imbalanced, we do not suggest preprocessing. (4) In wrapper-based feature selection, complicated searching method may not get better results, for example, genetic searching performs worse than best-first searching

C. Classifier Ensemble Design for Imbalanced Data Classification: A Hybrid Approach

Yubin Park et. al.2 presented two types of decision tree ensembles to handle issues faced by imbalanced data set. They have used an ensemble framework of α tree which gives higher AUROC values over different skewed data sets. A new splitting criterion has been introduced which uses diversification factor alpha (α).There is a need to focus on the diversity of base classifiers.

Mikel Galar et. al.3 proposed a novel ensemble construction technique known as EUSBoost that is based on RUBoost technique and makes combined use of evolutionary under-sampling with boosting. Evolutionary undersampling initially applies the random under-sampling technique to the skewed data sets, which are evolved until there are no further improvements in currently best under-sampled data set. Use of evolutionary under-sampling method has shown improvement in the performance of the base classifier.
Classifier ensemble can be formed by using different techniques which can be categorized into following categories.8:
1. Using different training sets: Introduce diversity by partitioning training dataset into N subsets and training an individual classifier with different subsets.
2. Using different feature subsets: Introduce diversity by training an individual classifier with a different subset of features.
3. Using different classifier models: Introduce diversity by combining different individual classifier.
4. Using different combination schemes: Introduce diversity by using different combination schemes.

The proposed work is carried out in two phases:
1. Re-sampling of the imbalanced data 2. Classifier Ensemble formation
Steps of the hybrid approach
1. Re-sample the imbalanced data
a. Over-sampling using Synthetic Minority Over-sampling Technique (SMOTE) SMOTE9, a well known over-sampling technique is applied to the imbalanced dataset in order to increase the number of samples of a minority class. This will help to reduce the imbalance ratio of the data set.
b. Under-sampling using modified random under-sampling technique Random under-sampling technique randomly selects and deletes some instances of majority class in order to decrease the imbalance ratio of the data set. But this may remove some necessary instances of the data set. To overcome this limitation, we have modified the algorithm by initially identifying the necessary data of the majority class and then applying random under-sampling to the remaining data.

2. Classifier Ensemble formation
a. Using different training set Bagging classifier ensemble is constructed using different training data subsets known as bootstrap samples. J48 is used as a base classifier and voting method is used to combine the predictions of individual classifiers.
b. Using different Classifier Model StackingC classifier ensemble is formed by using different classifier model. In our work, we have used three classifier models as base classifiers namely J48, LogisticRegression, and Bagging. Here bagging works as individual classifier but bagging itself is a classifier ensemble that has been formed by using different training subsets. Thus a hybrid approach is used to form the classifier ensemble.

Conclusion

An approach presented in this paper is to enhance the performance of classifier ensemble for imbalanced data. First concern is to modify random under-sampling approach in order to overcome its limitation of removing necessary instances of the majority class. For this, the necessary data is identified first and then the random under-sampling technique is applied on the remaining data. Second concern is to design a hybrid approach for constructing the classifier ensemble in which ensembles are formed using two methods i.e. using different training data sets and different learning models. Experimental results of proposed approach are compared with results of classifier ensemble that is constructed using different training data sets only. The evaluation results indicate that the presented approach outperforms the other technique in terms of AUC.

D. A Review on Ensembles for the Class Imbalance Problem: Bagging-, Boosting-, and Hybrid-Based Approaches

In this paper, Mikel Galar, Alberto Fernandez, Edurne Barrenechea, Humberto Bustince reviewed the state of the art on ensemble methodologies to deal with class imbalance problem. This issue hinders the performance of standard classifier learning algorithms that assume relatively balanced class distributions, and classic ensemble learning algorithms are not an exception. In recent years, several methodologies integrating solutions to enhance the induced classifiers in the presence of class imbalance by the usage of ensemble learning algorithms have been presented. However, there was a lack of framework where each one of them could be classified; for this reason, a taxonomy where they can be placed has been presented. We divided these methods into four families depending on their base ensemble learning algorithm and the way in which they address the class imbalance problem.

In this paper, we focus on two-class imbalanced datasets, where there is a positive(minority)class, with the lowest number of instances, and a negative (majority) class, with the highest number of instances. We also consider the imbalance ratio (IR) [54], defined as the number of negative class examples that are divided by the number of positive class examples, to organize the different data-sets.

In this paper, the AUC test results for all the algorithms in all datasets is presented. The results for non-ensembles and classic ensembles are shown. The test results for cost-sensitive boosting, boosting-based and hybrid ensembles, also the test results for bagging-based ones are shown. The results are shown in ascending order of the IR.

Traditionally, the accuracy rate (1) has been the most commonly used empirical measure. However, in the framework of imbalanced data-sets, accuracy is no longer a proper measure, since it does not distinguish between the numbers of correctly classified examples of different classes. Hence, it may lead to erroneous conclusions,i.e., a classifier that achieves an accuracy of 90% in a dataset with an IR value of 9, is not accurate if it classifies all examples as negatives.

$$Acc = \frac{TP + TN}{TP + FN + FP + TN} \qquad (1)$$

Conclusion

In this paper, the state of the art on ensemble methodologies to deal with class imbalance problem has been reviewed. This issue hinders the performance of standard classifier learning algorithms that assume relatively balanced class distributions, and classic ensemble learning algorithms are not an exception. In recent years, several methodologies integrating solutions to enhance the induced classifiers in the presence of class imbalance by the usage of ensemble learning algorithms have been presented.

III. CONCLUSION

We are trying to propose a new algorithm, Imbalanced Ensemble Feature Selection algorithm with sampled data using SMOTE, which will be designed to solve the low accuracy problem of minority classes when on an imbalanced class dataset classifications are executed. As the diversity metrics, we choose KW variance, and a punishment-reward factor was introduced to dynamically adjust the effect of the feature selections in order to increase the minority classes' accuracy. Different datasets and algorithm parameters will also produce different experimental results. We would compare IEFS with other three classification algorithms. From the result of the experiments, we may find that the predicting accuracy of minority class can be improved by the introduction of the attentions into diversity measure. In the future, to improve our algorithm we will try to use other searching methods.

REFERENCES

[1] Y. Mu, W. Ding, and D. Tao. Local discriminative distance metrics ensemble learning. Pattern Recognition, 46(8):2337–2349, 2013.

[2] H. Yin and K. Gai. An empirical study on preprocessing high dimensional class-imbalanced data for classification. In 2015 IEEE 17th International Conference on High-Performance Computing and Communications; The IEEE International Symposium on Big Data Security on Cloud pages 1314–1319, New York, USA, 2015.

[3] H. Yin and Y. Hu. An imbalanced feature selection algorithm based on the random forest. Journal of Sun Yat-sen University (Natural Science Edition), 4(5):59–65, 2014.

[4] M. Qiu, M. Zhong, J. Li, K. Gai, and Z. Zong. Phase-change memory optimization for the green cloud with a genetic algorithm. IEEE Transactions on Computers, 64(12):3528 – 3540, 2015.

[5] A. Holzinger and I. Jurisica. Knowledge discovery and data mining in biomedical informatics: The future is in integrative, interactive machine learning solutions. In Interactive Knowledge Discovery and Data Mining in Biomedical Informatics, pages 1–18. Springer, 2014.

[6] Yubin Park and Ghosh, J., "Ensembles of α-Trees for Imbalanced Classification Problems,", IEEE Transactions on Knowledge and Data Engineering, vol.26, no.1, pp.131-143, January 2014.

[7] Chawla, Nitesh V., et al. "SMOTE: synthetic minority over-sampling technique." Journal of artificial intelligence research (2002): 321-357.

[8] Yang, Pengyi, et al. "Sample subset optimization techniques for imbalanced and ensemble learning problems in bioinformatics applications." Cybernetics, IEEE Transactions on 44.3 (2014): 445-455.

[9] H. Yin and Y. Hu. A cost-sensitive algorithm based on the random forest. Engineering Journal of Wuhan University, 47(5):707–711, 2014.

[10] Mikel Galar, Alberto Fernandez, Edurne Barrenechea, Humberto Bustince ', A Review on Ensembles for the Class Imbalance Problem: Bagging-, Boosting-, and Hybrid-Based Approaches, July 2012.