A New Data Balancing Method for Classifying Multi-Class Imbalanced Genomic Data

Dewan Md. Farid DEWAN.MD.FARID@VUB.AC.BE
Computational Modeling Lab, Department of Computer Science, Vrije Universiteit Brussel
Pleinlaan 2, 1050 Brussels, Belgium

Ann Nowe & Bernard Manderick ANN.NOWE@VUB.AC.BE, BMANDERI@VUB.AC.BE
Computational Modeling Lab, Department of Computer Science, Vrije Universiteit Brussel
Pleinlaan 2, 1050 Brussels, Belgium

Keywords: clustering, imbalanced data, minority-majority classes

Abstract

Classification of multi-class imbalanced genomic data is a difficult task, as genomic data are noisy, high dimensional, small sample size that results overfitting and overlapping of classes. Traditional machine learning algorithms are very successful with classifying majority class instances compare to the minority class instances. The conventional data balancing methods alter the original data distribution, so they might suffer from overfitting or drop some potential information. In this paper, we propose a new method for dealing with multi-class imbalanced data based on clustering and selecting most informative instances from the majority classes.

1. Introduction

In recent years, classifying multi-class imbalanced data has become one of the key research topics in machine learning for data mining applications (Beyan & Fisher, 2015). Traditional machine learning algorithms successfully classify majority class instances, but misclassify the minority class instances in many high-dimensional data sets (Tomašev & Mladenić, 2013). Generally, sampling methods: (1) under-sampling and (2) over-sampling, cost-sensitive learning methods, and ensemble methods: (1) bagging and (2) boosting have been applied to handle the class imbalance problems (He & Garcia, 2009). To make the class distribution balanced, the under-sampling methods remove the majority class instances from the data set, while the over-sampling methods artificially generate and add minority class instances to the data set. So, over-sampling may increase the likelihood of overfitting and under-sampling may drop some potentially useful data/ information (Sun et al., 2015). The classification results are not stable in cost-sensitive learning methods, as it is difficult to get the accurate misclassification cost and different misclassification cost might result in different inductions. Also, the bagging and boosting use sampling methods to obtain balanced data in each iteration, so they might suffer from overfitting or drop some useful information (Galar et al., 2012).

This paper presents a clustering based multi-class imbalance data handling method that does not suffer from overfitting or drop some potential information. The proposed method converts the imbalanced data set into multiple balanced data sets and trains classifiers on each data set. It clusters the majority class instances into several clusters/ groups and finds the most informative instances in each cluster. Finally, several balanced data sets are created using these clusters with most informative instances, where the number of minority-majority classes instances are equal. In this paper, the proposed method is tested with the conventional class imbalance data handling methods: random under-sampling, random over-sampling, bagging, and boosting on genomic data (148 Exome data sets) of Brugada syndrome from the Centre of Medical Genetics, VUB UZ Brussel. Experimental analysis show that proposed method is more effective than the conventional class imbalance data handling methods.

Appearing in *Proceedings of Benelearn 2016*. Copyright 2016 by the author(s)/owner(s).

2. Proposed Data Balancing Method

Initially, we cluster the majority class instances into several clusters and find the most informative instances in each cluster. The informative instances are close to the center of cluster and border of cluster. Then several data sets are created using these clusters with most informative instances by combining the instances of minority classes. Every data set should have almost equal number of minority-majority classes instances. Finally, multiple classifiers are trained using these data sets. The voting technique is used to classify the existing/ new instances. In this paper, we use decision tree (C4.5) induction algorithm as a base classifier. The proposed data balancing method is illustrated in Fig. 1.

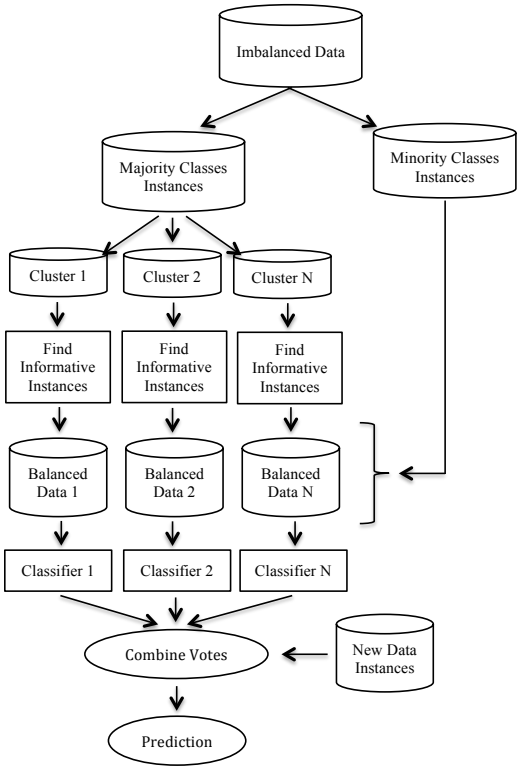


Figure 1. Proposed method for handling multi-class imbalanced data.

3. Experiments

We have tested the performance of data balancing methods using area under the ROC (Receiver Operating Characteristic) curve (AUC) on 2143 variants of Brugada syndrome (BrS) of 148 Exome data sets. BrS is a genetic disease, which increases the risk of sudden cardiac death (SCD) at a young age. Table 1 shows the average AUC values of 148 imbalanced Exome data sets for different imbalance data handling methods.

Table 1. Average AUC values of 148 imbalanced Exome data sets for different imbalance data handling methods.

Algorithm	Average AUC value
Random Under-Sampling	0.8923
Random Over-Sampling	0.8673
Bagging	0.8915
Boosting	0.9136
Proposed Method	0.9317

4. Conclusion

In this paper, we have presented a clustering based method for dealing with the multi-class imbalance problem. The proposed method does not remove majority class instances from the data set and add artificial minority class instances into the data set. Also, it does not change the class distribution. The proposed method performs more effectively than the conventional data balancing methods. In future, we will apply this method in other data mining applications.

Acknowledgments

The support for this research received from the BRiDGEIris (BRussels big Data platform for sharing and discovery in clinical GEnomics) project and FWO research project G004414N "Machine Learning for Data Mining Applications in Cancer Genomics".

References

Beyan, C., & Fisher, R. (2015). Classifying imbalanced data sets using similarity based hierarchical decomposition. *Pattern Recognition*, *48*, 1653–1672.

Galar, M., Fernandez, A., Barrenechea, E., Bustince, H., & Herrera, F. (2012). A review on ensembles for the class imbalance problem: Bagging-, boosting-, and hybrid-based approaches. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, *42*, 463–484.

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, *21*, 1263–1284.

Sun, Z., Song, Q., Zhu, X., Sun, H., Xu, B., & Zhou, Y. (2015). A novel ensemble method for classifying imbalanced data. *Pattern Recognition*, *48*, 1623–1637.

Tomašev, N., & Mladenić, D. (2013). Class imbalance and the curse of minority hubs. *Knowledge-Based Systems*, *53*, 157–172.