New Combining of SMOT and Tomek Links Sampling Methods for Functional Association Genes Prediction on Highly Imbalanced Data Sets

S. Soltani[a*], M. Jalali[a], J. Sadri[b]
a Department of Computer Engineering, Islamic Azad University, Mashhad, Iran
b Department of Computer Science & Software Engineering, Concordia University, Montreal, Quebec, Canada
*sima.soltani7@gmail.com

Abstract: The real-world data analyses especially in biology and medicine encountered highly imbalanced data sets (when the ratio between classes is high). Standard classification algorithms tend to be biased towards the major class to achieve maximum accuracy and ignore the minority samples as a noise, whereas prediction of the minor class is the crucial and base intent. Several approaches for dealing with imbalance problem have been introduced [1], such as sampling techniques [2] that are useful for making a balance training data set. In this paper, a new hybrid model is introduced to soften the skew problem into two high imbalanced data sets. It is applied the Synthetic Minority Oversampling Technique (SMOT) [3] for generating new samples similar to minority class instances due to predict minority instances accurately. moreover, because of existence extremely imbalance ratio in the data sets and necessity selecting appropriate samples from majority group, is used tomek links [1] undersampling method for removing majority instances. Appling this method can eliminate borderline samples from majority class. The feature vector is introduced for any gene in a data sets has 898 features, so is applied feature selection methods to achieve accurate prediction and increase classification performance. With due attention to results obtained in [4], we applied two feature ranking methods, Maximum mutual information (Mmi) and Maximum Pearson correlation coefficient (Mpcc). To improve the accuracy of genes prediction, a gene identification method is proposed through multiple classifier integrations. In the first stage, we made some balance learning sets by resampling of preprocessed data sets. In the second stage, the results of two classifiers, Decision tree, and k Nearest Neighbors, are combined with majority voting. The aim of this research is to predict 27 functional association genes which are related to resistance to virus converted, from 35445 unrelated genes. Rows of dataset indicate gene regions in the human genome and each gene region is introduced by features vector concerned to the promoter region of the gene.The data set is high dimensions and very high imbalance, therefore applying the most proposed methods related to imbalance problem in Literature individually tis not suitable for getting an acceptable prediction. According to the obtained results, feature selection and sampling methods improves accuracy and reduces training time. Voting based classifier reduces the classification errors. The proposed gene classification provides accuracy, sensitivity, and specificity at levels of 99.26%, 67%, and 99.28%, respectively. In compared to the proposed method presented in [4], the proposed method in this paper improve the prediction results.

Keywords: Gene prediction; SMOT; Tomek Links; Imbalanced Data Sets

References

[1] S. Kotsiantis, D. Kanellopoulos and P. Pintelas, “Handling imbalanced datasets: A review,” GESTS International Transactions on Computer Science and Engineering, 1.30 (2006), 25-36.

[2] E. Burnaev, P. Erofeev and A. Papanov, “Influence of resampling on accuracy of imbalanced classification,” In Eighth International Conference on Machine Vision (ICMV 2015) 2015

[3] N.V. Chawla, K.W. Bowyer, L.O. Hall and W.P. Kegelmeyer, “SMOTE: synthetic minority over-sampling technique,” Journal of artificial intelligence research 16 (2002), 321-357.

[4] S. Soltani, J. Sadri, H. Ahmadi Torshizi, “Feature selection and ensemble hierarchical cluster-based under-sampling approach for extremely imbalanced datasets,” Proceedings of the first International Conference on Computer and Knowledge Engineering, (ICCKE2011), 2011.