Combining Boosting and Active Learning for Mining Multi-Class 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: active learning, boosting, clustering, unlabeled data

Abstract

Boosting and active learning achieve high classification rate in many real world machine learning for data mining applications. This paper presents an optimal ensemble learning to improve the prediction accuracy of multi-class DNA variant classification employing boosting and active learning. We use naïve Bayes classifier and clustering to find the most informative unlabeled DNA variants as part of active learning and use boosting as a base classifier. The strategy of combining boosting and active learning is evaluated based on genomic data (148 Exome data sets) of Brugada syndrome from the Centre of Medical Genetics, VUB UZ Brussel.

1. Introduction

Labeling of unlabeled DNA variants is expensive and time-consuming. Active learning uses a user/ oracle to label the unlabeled data on demand and tries to improve the classification accuracy by posing minimum queries to the user (Lughofer, 2012). It selects the most informative instances from unlabeled data that the knowledge of their corresponding labels improves the accuracy of learning algorithms. Several active learning methods have been developed in the last decade (Settles, 2010). Ghasemi et al. (2011) proposed a method for uncertainty sampling using Bayes' rule and density estimation in which only one class instances and unlabeled data were available. Yang et al. (2015) proposed an uncertainty sampling with diversity maximization for multi-class active learning that evaluated all the active pool instances to train a model by making the uncertainty evaluation more accurate.

In this paper, we combine boosting and active learning for classifying multi-class genomic data to improve the prediction accuracy of DNA variants of Brugada syndrome (BrS). BrS is a genetic disease, which increases the risk of sudden cardiac death (SCD) at a young age. We use naïve Bayes classifier and clustering to find the most informative instances for labeling and use boosting algorithm as a base classifier to train the model. The performance of proposed method tested on genomic data (148 Exome data sets) of BrS from the Centre of Medical Genetics, VUB UZ Brussel, Belgium. Experimental analysis show promising results compared to other similar ensemble learning methods.

2. Active Learning

Active learning achieves high accuracy using the number of instances to learn a concept can often be much lower than the number required in typical supervised learning. It interactively queries a user/ expert for class labels of unlabeled instances. The objective is to train a classifier using as few labeled instances as possible by selecting the most informative instances. Let the data, D contains both set of labeled data, D_L and set of unlabeled data, D_U. Initially, a model, M^* trains using D_L. Then a *querying function* uses to select unlabeled instances, $X_U \in D_U$ and requests a user for labeling, $X_U \rightarrow X_L$. After X_L is added to D_L and train M^* again. The process repeats until the user is satisfied. In active learning, the research

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

challenge is to find the best method for selecting the unlabeled instances, $X_U \in D_U$. Active learning model is illustrated in Fig. 1.

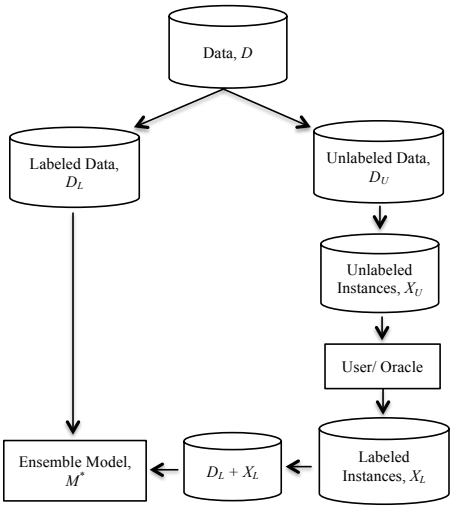


Figure 1. Diagram illustrating the active learning process.

3. Proposed Method

The naïve Bayes (NB) classifier and clustering are used to find the most informative instances for labeling as part of active learning. Adaptive boosting (AdaBoost algorithm) with NB classifier is used as base classifier. The unlabeled instances are selected for labeling using the following two strategies:

- Instances close to centers of clusters and borders of clusters.
- If the posterior probabilities of instances are equal/ very close.

4. Experiments

We have tested the performance of ensemble methods on 2143 variants of BrS from 148 Exome data sets. Table 1 and Table 2 summarised the data sets. Table 3 shows the accuracy and F-score of different ensemble methods on training data.

Table 1. Classification of DNA variants for BrS.

Class	Label
Class I	Nonpathogenic
Class II	VUS1 - Unlikely pathogenic
Class III	VUS2 - Unclear
Class IV	VUS3 - Likely pathogenic
Class V	Pathogenic

Table 2. Number of variants in 148 Exome data sets.

	VCF File	Gene Panel	BrS
Total	19687	17795	2143
Average	133	120	14
Maximum	267	228	135
Minimum	6	6	0

Table 3. The accuracy and F-score of ensemble methods.

Algorithm	Classification accuracy (%)	F-score (weighted avg.)
Random Forest	92.3	0.93
Bagging	87.5	0.83
Boosting	91.66	0.9
Proposed Method	94.73	0.93

5. Conclusion

Combining boosting and active learning is a new strategy in supervised learning, which achieves high accuracy for classifying multi-class genomic data. This approach can be applied 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

Ghasemi, A., Rabiee, H. R., Fadaee, M., Manzuri, M. T., & Rohban, M. H. (2011). Active learning from positive and unlabeled data. *IEEE 11th International Conference on Data Mining Workshops*, 244–250.

Lughofer, E. (2012). Hybrid active learning for reducing the annotation effort of operators in classification systems. *Pattern Recognition*, *45*, 884–896.

Settles, B. (2010). *Active learning literature survey* (Technical Report 1648). Computer Science, University of Wisconsin - Madison.

Yang, Y., Ma, Z., Nie, F., Chang, X., & Hauptmann, A. G. (2015). Multi-class active learning by uncertainty sampling with diversity maximization. *International Journal of Computer Vision*, *113*, 113–127.