Effect of Classifiers in Consensus Feature Ranking for Biomedical Datasets

Shobeir Fakhraei[1,2]
shobeir@wayne.edu

Hamid Soltanian-Zadeh[2,3]
hszadeh@ut.ac.ir

Farshad Fotouhi[1]
fotouhi@wayne.edu

Kost Elisevich[4]
nskoe@neuro.hfh.edu

1. Dept. of Computer Science
Wayne State University
Detroit, MI, USA

2. Image Analysis Lab.
Dept. of Radiology
Henry Ford Health System
Detroit, MI, USA

3. CIPCE, School of Elec. and Comp. Eng.
University of Tehran
Tehran, Iran

4. Dept. of Neurosurgery
Henry Ford Health System
Detroit, MI, USA

ABSTRACT
Many informative aspects of medical datasets may be extracted from comparative study of features discriminative power. Recently, consensus feature rankings have been proposed to achieve robust, unbiased and reliable rankings of attributes. We have studied the effect of classifier inclusion in a consensus feature ranking method for a medical dataset with missing values and class imbalanced data. Ability of consensus feature rankings to demonstrate superior performance with unseen classifiers is also studied in this paper.

Categories and Subject Descriptors
I.5.2 [**Pattern Recognition**]: Design Methodology-feature evaluation and selection

General Terms
Algorithms, Performance, Experimentation

Keywords
Feature ranking; consensus classification; classifier ensemble; missing value; class imbalance

1. INTRODUCTION
"Feature selection" and "feature ranking" ease data understanding and reduce measurement and storage requirements. Feature selection is utilized in biomedicine and bioinformatics in many applications such as diagnostic evaluation of medical tests and discovery of biomarkers [1]. Recently, "ensemble methods" have been used in feature ranking and selection to mitigate the problems of traditional methods such as poor accuracy, bias, and stability [1]. When multiple classifiers are used in this combination, the method is referred to as "consensus" feature selection or ranking. Studies like [3] have shown that this approach improves the performance of the methods.

When applying these methods on medical datasets, one has to consider the class imbalanced data and missing values as a common problem. Considering such characteristics, we have studied the effect of specific classifiers in consensus feature ranking. The study is conducted on clinical data of patients with temporal lobe epilepsy and their surgical results extracted from human brain image database system (HBIDS) [4].

2. METHOD
Five of the most commonly used classifiers in biomedicine are included in this study both in the ranking and evaluation phases. Their effect in consensus feature ranking, evaluated by themselves and by other classifiers, is studied. To avoid negatively affecting the reliability of the model, we did not estimate the missing values and performed the study only based on properly recorded values. Thus, certain parts of the dataset were eliminated, with adverse effect on data distribution. We used the area under receiver operating characteristic (ROC) curve (AUC) as a performance evaluator for individual features, to handle the balance problem.

To measure the individual effect of five classifiers in consensus feature ranking, features are ranked six times based on different consensus rankings; once based on the fusion of scores from all the classifiers, and five times based on fusion of scores from all classifiers excluding one at a time. These six consensus rankings are then evaluated by all of the classifiers to observe the effect of each classifier in the consensus ranking.

To conduct the study, featured are individually evaluated with a single classifier and scored based on its classification performance. Performance of each feature is measured by its average AUC based on the leave-one-out technique. The instances that had a missing value in the considered feature are eliminated from the dataset. To rank the features based on consensus rankings, the AUC from several classifiers are combined into a single consensus score using "median" as the fusion function. The features are then sorted and ranked based on this consensus scoring.

2.1 Evaluation Technique
To evaluate the feature rankings, α features from the top of the ranked features were selected and the predictive power of this feature subset was tested with a classifier via cross validation [2]. We eliminated the samples with missing values in the evaluations phase. To use the maximum possible instances for each feature subset, we used the samples that have all the values for only the features in the subset. In such a case, the number of instances varies for each feature subset, making the comparison of the ranking methods with different feature subsets difficult. To address the above problem, we used a performance index (*PI*) which is computed by Equation (1).

$$PI(n,c) = \sum_{i=1}^{n}\left(\frac{F_i\_ins}{i}.AUC\left(c(F_i)\right)\right)/\sum_{i=1}^{n}\left(\frac{F_i\_ins}{i}\right) \quad (1)$$

where n is the number of features considered in the calculation and c is the evaluating classifier. F_i is the set of i features with the highest fusion score and F_i_ins is the numbers of instances that have all the values for features in F_i. $AUC\left(c(F_i)\right)$ represents the average *AUC* for evaluation of F_i on c, using the leave-one-out technique.

A consideration in this formula is that the ranking methods that achieve a higher accuracy with fewer features and more instances are preferable. For this reason, the number of features appears in

Copyright is held by the author/owner(s).
DTMBIO'10, October 26, 2010, Toronto, Ontario, Canada.
ACM 978-1-4503-0382-8/10/10.

the weight factor as $1/i$ and the number of instances as F_i_ins.

3. EXPERIMENTAL RESULTS

The dataset used in the following experiments is from HBIDS [4] which contains medical data of epilepsy patients. The main task in this dataset is a binary classification that predicts the patients' lateralization (side of abnormality). The database contains 197 medical features and 145 patients. Five classifiers namely naïve Bayes, support vector machine, 3-nearest neighbors, bagging, and logistic regression are used in this study. $PI(n, c)$ of the fusion rankings are calculated for $n \in \{1, \ldots, 19\}$. In some subsets with more than nineteen features, evaluating with cross validation was not possible due to inadequate number of instances without missing values. The results are plotted in the charts presented in Figure 1. Points in the charts correspond to n in $PI(n, c)$.

As shown in Figure 1a rankings that exclude bagging and k-nearest neighbors (KNN) outperform others, indicating negative effects of these two classifiers on consensus ranking for SVM. Interestingly SVM itself has neutral effect on the performance of fusion ranking. It means that consensus ranking based on other classifiers contains adequate information to perform well with SVM. Removal of logistic regression and naïve Bayes from the fusion has adversely affected the performance. Charts in Figures 1b and 1d are the results of the evaluation on logistic regression and naïve Bayes, respectively. All of the rankings perform similarly on these classifiers showing that they are not highly sensitive to feature selection. However, it is also noticeable that removal of these classifiers themselves from the fusion has not negatively affected the results.

Results based on evaluation with the KNN classifier are shown in Figure 1c. Exclusion of the naïve Bayes and logistic regression demonstrate a positive effect on performance, while removal of the others including the KNN itself does not have a significant impact on the results. Feature rankings have also been tested with bagging. As shown in Figure 1e, naïve Bayes and logistic regression play a negative role in the combination. It is also notable that removal of the bagging classifier itself does not affect the performance of the rankings significantly.

4. DISCUSSION AND CONCLUSION

In these studies, we applied the consensus feature ranking to medical datasets with many missing values and imbalanced data. We demonstrated the effect of each classifier in consensus features ranking. Most importantly, it has been shown that the performance of the consensus feature ranking on a classifier is not highly dependent on inclusion of that classifier itself in the fusion. This indicates that features ranked based on fusion of scores from multiple classifiers perform well on unseen classifiers. This ranking plays an important role in data-warehousing, where data are gathered with the possibility to be used with new emerging classifiers in the future. In the continuation of this work, other classifiers and fusion functions can be studied and evaluated to achieve better understanding of the matter.

5. ACKNOWLEDGMENTS

This work was supported in part by NIH grant R01-EB002450.

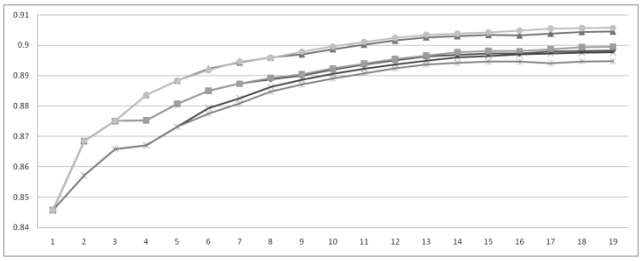
6. REFERENCES

[1] Y. Saeys, I. Inza, P. Larranaga, "A review of feature selection techniques in bioinformatics," Bioinformatics, vol. 23, p. 2507, 2007.

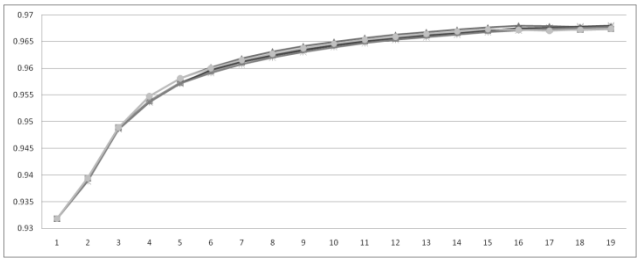
[2] I. Guyon, A. Elisseeff, "An introduction to variable and feature selection," Journal of Machine Learning Research, vol. 3, pp. 1157-82, 2003.

[3] J. Dutkowski, A. Gambin, "On consensus biomarker selection," BMC bioinformatics, vol. 8, p. S5, 2007.

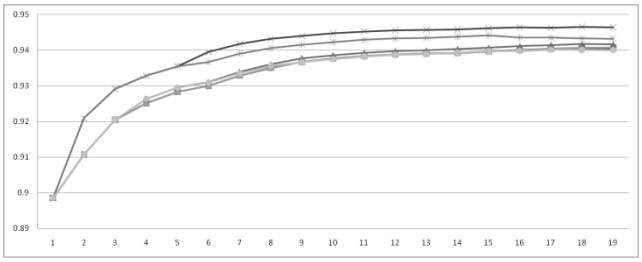
[4] M.R. Siadat, H. Soltanian-Zadeh, F. Fotouhi, K.Elisevich, "Content-based image database system for epilepsy," Computer Methods and Programs in Biomedicine, vol. 79, pp. 209-226, 2005.



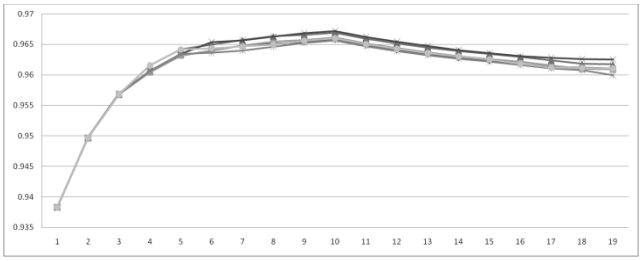
(a) Evaluation on SVM



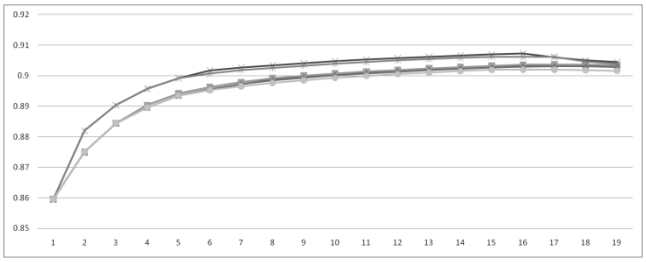
(b) Evaluation on Logistic-Regression



(c) Evaluation on K-Nearest-Neighbors



(d) Evaluation on Naïve-Bayes



(e) Evaluation on Bagging

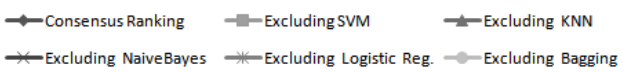


Figure 1. Effect of each classifier on consensus feature rankings evaluated on five different models.