CONSENSUS FEATURE RANKING IN DATASETS WITH MISSING VALUES

SHOBEIR FAKHRAEI[1,2], HAMID SOLTANIAN-ZADEH[2,3], FARSHAD FOTOUHI[1], KOST ELISEVICH[4]

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

Development of a feature ranking method based upon the discriminative power of features and unbiased towards classifiers is of interest. We have studied a consensus feature ranking method, based on multiple classifiers, and have shown its superiority to well known statistical ranking methods. In a target environment such as a medical dataset, missing values and an unbalanced distribution of data must be taken into consideration in the ranking and evaluation phases in order to legitimately apply a feature ranking method. In a comparison study, a Performance Index (PI) is proposed that takes into account both the number of features and the number of samples involved in the classification.

INTRODUCTION

It is known that the prediction accuracy of practical machine learning algorithms degrades when faced with many features that are not necessary for predicting the desired output. "Feature selection", the removal of irrelevant features in a dataset, not only circumvents the curse of dimensionality but also makes the learning process faster and the model simpler. It also facilitates data visualization and data understanding while reducing measurement and storage requirements.

Another aspect of feature selection is achieving a better understanding of the data important to particular domains such as medicine. Discovering which medical tests have higher diagnostic value than the others is valuable. In such domains, the accuracy of a classifier is also important. A high number of false negatives might deprive some patients from the required attention, while a high false positive rate will cause unnecessary concern and a waste of medical resources.

A closely related concept to feature selection is "feature ranking", which is sometimes regarded as a relaxed feature selection method. Feature ranking involves the sorting of features according to a "feature quality index" that reflects the relevance, information, or discriminating capability of the feature.

Imprecise results, computational complexity and overfitting of a feature subset to a specific classifier have prompted new approaches that use modifications of ensemble methods and consensus decisions for feature ranking. In most consensus methods, statistical measures are combined. In the ensemble methods, a single classifier is used to evaluate the performance of a feature. This again either does not utilize the power of classifiers to find features with the highest classification accuracy or causes the ranking results to be biased towards a specific classifier. In this paper, we combine the results from multiple classifiers to mitigate such problems.

We have studied five of the best known classifiers and applied the method to rank medical features in a clinical database with missing values and class-imbalanced data. The main question addressed in this paper regards establishing whether consensus feature ranking outperforms traditional methods and whether it would be unbiased towards classifiers in an environment with missing values and unbalanced distribution.

THE PROPOSED METHOD

1. Framework

In our method, each feature is individually assessed with a single classifier and scored based on its classification performance. In order to avoid fabrication of data instances, prior to applying a classifier on the data, the instances that had a missing value in the considered feature are eliminated from the dataset.

The scores from several sources are combined into a single consensus score. The features are then sorted and ranked based on this consensus scoring. At the evaluation phase, feature subsets are formed by selecting α number of top-ranking features. The subsets are evaluated based on their classification accuracy using 10-fold cross validation with multiple classifiers and their performance index is calculated based on the results.

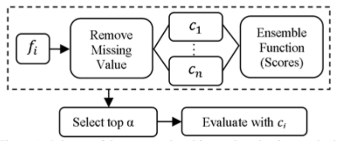


Figure 1. Schema of the proposed ranking and evaluation method.

2. Ranking Measure

We use multiple classifiers as a tool to perform the rankings. Since classification accuracy is sensitive to unbalanced distributions, we evaluate predictive power of each feature based on the area under the ROC curve (AUC).

3. Ensemble Function

In order to rank the features, we use the ranking scores from different ranking measures, combine them using an ensemble function and sort (rank) the features accordingly. Our preliminary studies show that superior performance is achieved when using the mean as the ensemble function. Therefore, in order not to complicate the study, we only consider the mean as the ensemble function.

4. Evaluation Technique

To handle the problem of many missing values without highly affecting the results, we eliminate the samples with missing values. In such a case, the number of instances varies for each feature subset. For example, the samples which have a value for two features might not have all the values for three features. To address this problem, we used a performance index (PI) which is computed by

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

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. And $AUC(c(F_i))$ represents the average AUC for evaluation of F_i with c, using the 10-fold cross validation technique.

A consideration in this formula is that the ranking methods that achieve a higher accuracy with fewer features and more instances are preferable.

EXPERIMENTAL RESULTS

The dataset used in the following experiments is from the Human Brain Image Database System (HBIDS) developed in the Radiology Department of Henry Ford Hospital. The dataset 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 146 patients.

We compare the ranking of the features from the consensus method with the rankings from the information gain and chi-square statistics ranking methods using the Performance Index (PI). The five classifiers used in these experiments are decision tree (DT), naïve Bayes (NB), support vector machines (SVM), k-nearest neighbors (KNN), and multilayer perceptron (MLP).

PI(n,c) of the consensus ranking, information gain and chi-square statistic are calculated were *n* is between one and eighteen. In some subsets with more than eighteen features, evaluating with 10-fold cross validation is not possible due to the number of instances being less than ten.

we have also included best and worst cases of the single classifier rankings in addition to the three ranking methods in Fig. 2 to demonstrate the performance of the consensus ranking method with respect to the minimum and maximum possible accuracies that could be achieved using the same number of features in a feature subset.

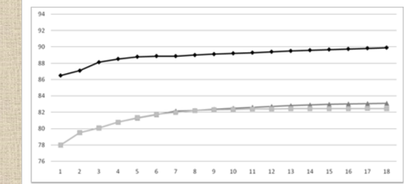
DISCUSSION AND CONCLUSION

In these studies, with a proposed weighted performance measure and classification accuracy, it has been shown that the consensus ranking method outperforms two commonly used ranking methods in data mining and machine learning. The minimum and maximum prediction accuracies of these methods along with single classifier ranking have also been presented.

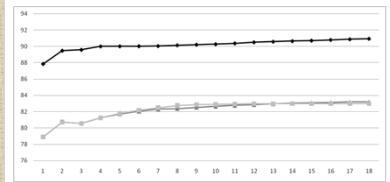
In general, the consensus ranking method prioritized the more informative features appropriately. In both the PI and accuracy charts, the current method provided more reliable results on subsets with small numbers of features. As a feature subset became more populated, classification accuracy remained at a level approximating that generated by other methods, indicating exclusion of completely irrelevant features in the studied portion.

The consensus ranking methods always performed consistently and no significant bias towards a single classifier was observed. However, the consensus ranking showed slightly better performance results when evaluated with NB and KNN classifiers. Evaluation with SVM and MLP demonstrated inferior results than the other two mentioned classifiers. The ranking performed worst with DT classifier.

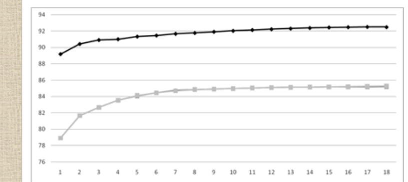
FIG1. PI OF THE RANKING METHODS



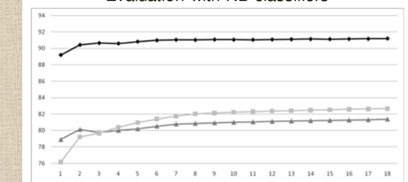
Evaluation with SVM classifiers



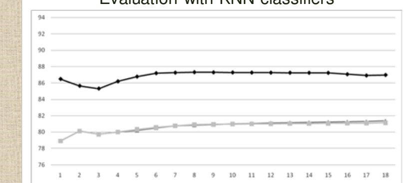
Evaluation with MLP classifiers



Evaluation with NB classifiers



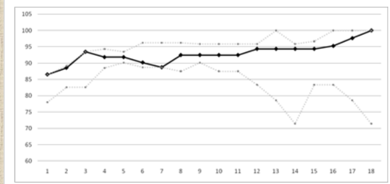
Evaluation with KNN classifiers



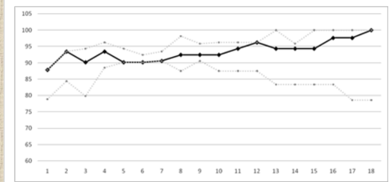
Evaluation with DT classifiers

Consensus Ranking — Information Gain — Chi-Square

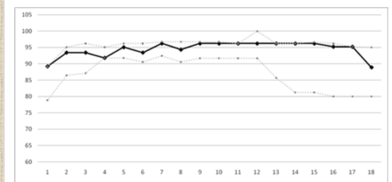
FIG2. CLASSIFICATION ACCURACY OF THE CONSENSUS RANKING METHOD



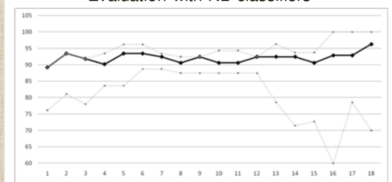
Evaluation with SVM classifiers



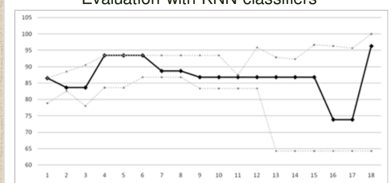
Evaluation with MLP classifiers



Evaluation with NB classifiers



Evaluation with KNN classifiers



Evaluation with DT classifiers

Consensus Ranking — MIN — MAX

REFERENCES

1. R. Kohavi and G. H. John, "Wrappers for feature subset selection," *Artificial Intelligence,* vol. 97, pp. 273-324, 1997.
2. K. Chrysostomou, et. al, "Combining multiple classifiers for wrapper feature selection," *Int. J. of Data Mining, Modelling & Management,* vol. 1, pp. 91-102, 2008.
3. R. E. Abdel-Aal, "GMDH-based feature ranking and selection for improved classification of medical data," *J. of Biomedical Informatics,* vol. 38, pp. 456-68, 2005.
4. 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.
5. M. Makrehchi and M. S. Kamel, "Combining feature ranking for text classification," in *2007 IEEE Int. Conf. on Systems, Man, and Cybernetics*, Montreal, QC, Canada, 2007, pp. 510-515.
6. M. Hall, et. al, "The WEKA data mining software: an update," *SIGKDD Explor. Newsl.,* vol. 11, pp. 10-18, 2009.
7. X.-W. Chen and M. Wasikowski, "FAST: A roc-based feature selection metric for small samples and imbalanced data classification problems," in *14th ACM KDD 2008,* Las Vegas, NV, United states, 2008, pp. 124-132.