2010 Ninth International Conference on Machine Learning and Applications

Consensus Feature Ranking in Datasets with Missing Values

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

Hamid Soltanian-Zadeh[2,3]
hamids@rad.hfh.edu

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— **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.**

Keywords-feature ranking; feature selection; consensus ranking; heterogeneous classifier ensemble; missing value; class imbalanced distribution;

I. 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 [1]. "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 [2].

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 [3].

Most feature ranking methods are based on statistical measures. Otherwise, the prediction accuracy of a feature is considered as a ranking score by using only a single classifier, similar to the wrapper approach [1]. 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 of patients with temporal lobe epilepsy and their surgical results called Human Brain Image Database System (HBIDS) [4]. Like many other medical datasets, HBIDS contains a large number of attributes and a relatively small number of data records [3]. Moreover, not all of the medical tests are performed for every patient, which leaves the data with many missing values. Another common problem with most medical datasets similar to HBIDS is the disproportionate representation of the target cohort and that of a comparative control population.

To mitigate the missing value problem in critical domains as in medical applications, one should not negatively affect accuracy and reliability of the classifier by fabricating and estimating the data. Therefore, to study the predictive power of a specific feature, we only use patients that have a value for that feature in their records and eliminate those patients without. The elimination of certain instances from the dataset may adversely affect data distribution. To tackle the unbalanced distribution problem, we have evaluated the feature performance based on the area under receiver operating characteristic (ROC) curve (AUC) instead of the classification accuracy.

The main question addressed in this paper regards

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

978-0-7695-4300-0/10 $26.00 © 2010 IEEE
DOI 10.1109/ICMLA.2010.117

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.

II. Related Works

Some of the methods related to consensus feature ranking are highlighted in this section. Certain methods focus on medical datasets which is our main target environment. However, the methods described below either do not utilize classifiers as ranking measures or, use only a single classifier for ranking features. They also do not consider missing values and unbalanced data distribution, or study the bias of the consensus method towards specific classifiers.

Group method of data handling (GMDH)-based feature ranking for medical data uses the GMDH learning algorithm to automatically select the optimal predicting features at different levels of user specified model complexity [3]. ROC is used to evaluate the classifier performance.

Makrehki and Kamel [5] combined feature rankings for text classification. The feature ranking measures were considered as voters and features as candidates. Two Borda techniques, the Fuzzy and Nash voting methods, are examined on multiple feature ranking measures. Final ranks are evaluated using a support vector machine (SVM). By comparing the minimum and maximum performance of single feature rankings to those of the combined ranking, it was shown that combined feature ranking offered a reliable result, independent of the voting schema.

Chrysostomou et al [2] used a wrapper-based decision tree (WDT) to overcome the problem of bias of traditional wrappers. The WDT combines multiple classifiers and uses a decision tree to visualize the relationships among the features. Four different classifier families are considered in this study. The wrapper and the 10-fold cross validation are applied in this method in the Waikato environment for knowledge analysis (WEKA) [6]. The score of a feature given by a classifier ranges from zero (the feature is not selected at all) to ten (the feature is always selected by the cross validation). The median function is used to combine the results of different classifiers. Performance of the selected feature subset is evaluated using a SVM classifier.

III. The Proposed Method

In this study, we assess the effectiveness of consensus feature ranking from two perspectives to establish whether this method performs better than the traditional ranking methods by comparing it with well known information gain and chi-square statistics feature rankings, and single classifier feature rankings. For these comparisons, we have proposed a performance index suitable for datasets with many missing values. The other goal is to examine the consensus feature ranking with several classifiers to observe any bias from consensus feature ranking towards any specific classifier. Five of the most widely used classifiers are included in this study both in ranking and evaluation phases. In this section, our method's overall framework, ranking measure, ensemble function, and evaluation technique are explained.

A. 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.

We use the notations summarized in Table I to describe the methods used in our framework. The framework is implemented using the WEKA [6] open source software codes in JAVA. The overall schema of our method is shown in Figure 1.

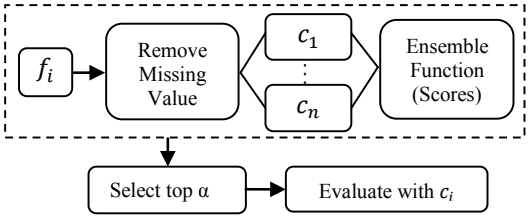


Figure 1. Schema of the proposed ranking and evaluation method.

TABLE I. Notations and Descriptions

Symbol	Description
F	Subset of features in the dataset containing $\{f_1, f_2, \dots, f_n\}$ where F may contain any number of features from all the features in the dataset to only one feature f_i
$\|F\|_{Ins}$	Number of all instances or samples in the dataset for feature set F
$\|F\|_{Att}$	Number of all attributes or features in the feature set F
C	Set of classifiers containing $\{c_1, c_2, \dots, c_n\}$ where $c_i(F)$ means applying the classifier c_i on the samples of feature set F
$NVR(F)$	Function that removes all the null values from F
$AUC(c_i(F))$	Area under the receiver operating characteristic (ROC) curve when applying the classifier c_i to the dataset only containing the features from F. ACC could be used instead of AUC to indicate accuracy
$TOP(F, X, n)$	Selected top n number of features from F based on the X criterion which may be accuracy, AUC, or a ranking methods

B. 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) [7]:

$$AUC\left(c_i\left(NVR(f_i)\right)\right) \quad (1)$$

C. 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. The ensemble function can be written as the following where *FS* is the ensemble score.

$$ES(f_k) = \frac{\sum_{c_i \in C} AUC\left(c_i(NVR(f_k))\right)}{\sum_{c_i \in C} 1} \quad (2)$$

D. Evaluation Technique

A common method used to evaluate feature ranking is to select α features from the top of the ranked features and test the predictive power of this feature subset with a classifier via cross validation. We measure the predictive accuracy of each subset using different classifiers and 10-fold cross validation. The evaluation can be written as (3) where α is the number of top ranking features included in the feature subset and φ is the ranking method.

$$ACC\left(c_i\left(NVR\left(TOP(F,\varphi,\alpha)\right)\right)\right) \quad (3)$$

The AUC can also be used instead of accuracy in (3). However, in our preliminary studies, a significant difference was not observed in the outcome of this phase when using either one. Therefore, we only report the accuracy results.

As mentioned earlier, to handle the problem of many missing values without highly affecting the results, we eliminate the samples with missing values. However, in our test environment, the samples that have all of the features are not many. To use the maximum possible instances for each feature subset, we use the samples that have all the values for only the features in the subset and not for all the features in the dataset.

In such a case, the number of instances varies for each feature subset. For example, the samples which have a value for both $\{f_1, f_2\}$ might not have all the values for $\{f_1, f_2, f_3\}$. This makes it hard to compare the ranking methods with different numbers of feature subsets. The situation is worse when considering that subsets with the same number of features might also have a different number of instances, since different feature ranking methods generate different feature subsets.

To tackle the above problem, we propose a merit that considers the number of features and the number of instances and calculate the overall performance of a feature ranking. The calculated value which we call performance index (*PI*) is computed by equation (4) which is the weighted average of the classification accuracies of the subsets containing only one feature to that of the subset containing *N* features.

$$PI(N,\varphi) = \frac{\sum_{i=1}^{N}\left(\frac{|F_{i\varphi}|_{Ins}}{|F_{i\varphi}|_{Att}} . ACC\left(c_i(F_{i\varphi})\right)\right)}{\sum_{i=1}^{N}\left(\frac{|F_{i\varphi}|_{Ins}}{|F_{i\varphi}|_{Att}}\right)} \quad (4)$$

where $F_{i\varphi} = NVR\left(TOP(F,\varphi,i)\right)$

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 the weight factor as $1/|F_{i\varphi}|_{Att}$ and the number of instances as $|F_{i\varphi}|_{Ins}$.

IV. 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 [4]. 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 formula (4). 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).

$PI(N,\varphi)$ of the consensus ranking, information gain and chi-square statistic are calculated for $N \in \{1,\ldots,18\}$. 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.

The results are plotted in the charts presented in Figure 2 and Figure 3. In Figure 2a, SVM is used as the φ in formula (4). The overall performance of the consensus ranking is better than the other methods while the information gain and chi-square statistics ranking perform similarly. Note that for the SVM classifier, the consensus ranking has gathered the most informative features at the top, where after the first few features, adding more features does not have much improvement effect. On the other hand, the PI for the information gain and chi-square increases at a higher rate towards the end. This means that there are informative features in the middle of the list.

Figure 3a shows the accuracy of different subsets containing the top α features from the consensus ranking method when evaluated with SVM using (3); α is varied from 1 to 18. In order to show the overall performance of the consensus ranking method from another perspective, two other guidelines are drawn in Figure 3. The line at the top and the bottom are the maximum and the minimum possible accuracies that could be achieved at a point, using different ranking methods. A point in the diagram corresponds to a number of features in the feature subset. For the minimum and maximum possible accuracies, we have also included the single classifier rankings in addition to the three ranking methods in Figure 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. The minimum and maximum values are formulated as:

$$F_\tau(\alpha, c_i) = \tau\left(ACC\left(c_i\left(NVR\left(TOP(F, \varphi, \alpha)\right)\right)\right)\right) \quad (5)$$

where φ ∈ {consensus, info-gain, chi-square, SVM, MLP, KNN, DT, NB} and τ ∈ {min, max}.

Although variation of the number of instances in subsets with the same number of features will affect the performance, Figure 3a demonstrates that the consensus ranking method generates near maximum classification accuracy, especially for the first few of the highly ranked features.

The same experiments are repeated using MLP as the evaluating classifier. The PIs in Figure 2b demonstrate that again the consensus ranking method outperforms the other two methods. The same behavior is observed with higher number of features. Figure 3b demonstrates the accuracy of the consensus feature ranking method evaluated by MLP. Note that the same consensus ranking evaluated with the SVM generated reasonably good overall performance with both classifiers.

In the next studies, the consensus ranking and other ranking methods are evaluated using NB and KNN (with K=3) classifiers; Figure 2c and 2d demonstrate the PI of the three ranking methods. Figure 3c and 3d show the overall accuracy of the consensus ranking method using NB and KNN classifiers. At most of the points, the consensus ranking method performed well with near maximum accuracy. It is interesting to note that in Figure 2d, although the chi-square statistics method did not have a good performance in ranking of the first feature, it outperformed the information gain method in the next ranked features.

The ranking methods are also evaluated via DT as the last classifier. The performance indices of the methods are shown in Figure 2e. Although the overall performance of the consensus ranking method is better than the other two, the margin between them is lower than the other classifiers. Another noticeable behavior of this classifier is the finding that the PIs of the consensus ranking method for the second and third point are lower than that of the first point. This means that the first feature is ranked well but the second and third features are not. The same behavior is shown in the accuracy chart presented in Figure 3e. The first feature is ranked perfectly and then some other related features are ranked from four to six. However, the rest of the ranked features do not contribute to the accuracy of the classifier. With regards to the minimum and maximum values, it is also notable that the overall performance of the DT classifier is not as good as the other classifiers used in our experiments.

V. 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 (not combined) 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.

References

[1] R. Kohavi and G. H. John, "Wrappers for feature subset selection," *Artificial Intelligence,* vol. 97, pp. 273-324, 1997.

[2] K. Chrysostomou, S. Chen, X. Liu, "Combining multiple classifiers for wrapper feature selection," *International Journal of Data Mining, Modelling and Management,* vol. 1, pp. 91-102, 2008.

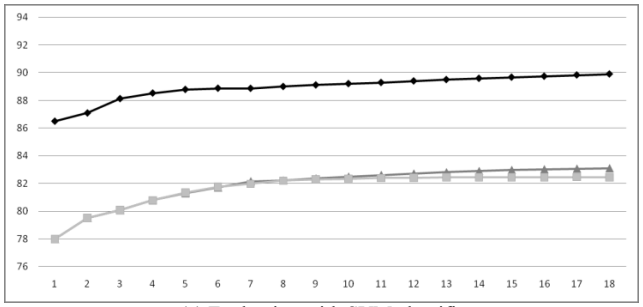
[3] R. E. Abdel-Aal, "GMDH-based feature ranking and selection for improved classification of medical data," *Journal of Biomedical Informatics,* vol. 38, pp. 456-68, 2005.

[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.

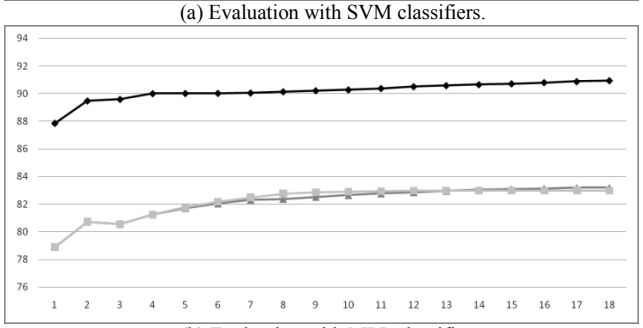
[5] M. Makrehchi and M. S. Kamel, "Combining feature ranking for text classification," in *2007 IEEE International Conference on Systems, Man, and Cybernetics*, Montreal, QC, Canada, 2007, pp. 510-515.

[6] M. Hall, E. Frank, G. Holmes,B. Pfahringer, P. Reutemann, I. H. Witten, "The WEKA data mining software: an update," *SIGKDD Explor. Newsl.,* vol. 11, pp. 10-18, 2009.

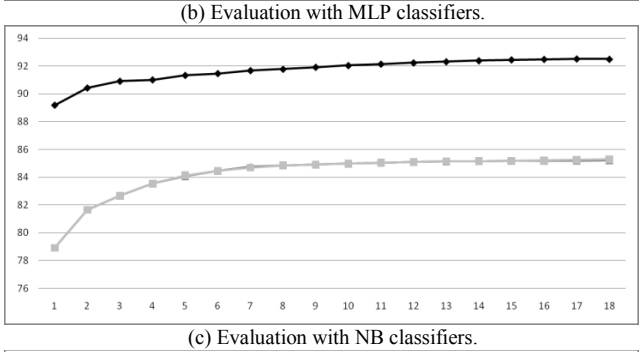
[7] X.-W. Chen and M. Wasikowski, "FAST: A roc-based feature selection metric for small samples and imbalanced data classification problems," in *14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2008,* Las Vegas, NV, United states, 2008, pp. 124-132.



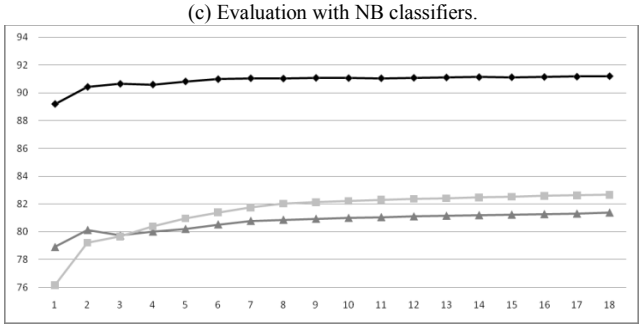
(a) Evaluation with SVM classifiers.



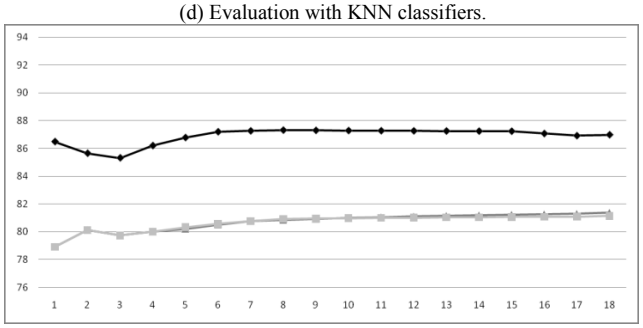
(b) Evaluation with MLP classifiers.



(c) Evaluation with NB classifiers.



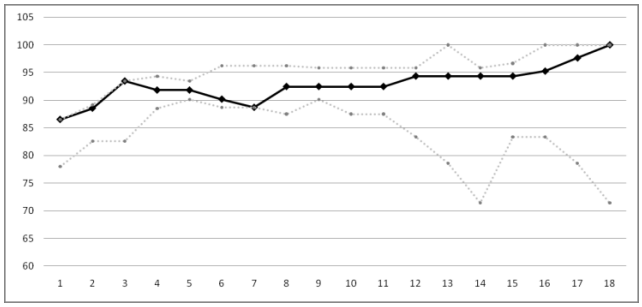
(d) Evaluation with KNN classifiers.



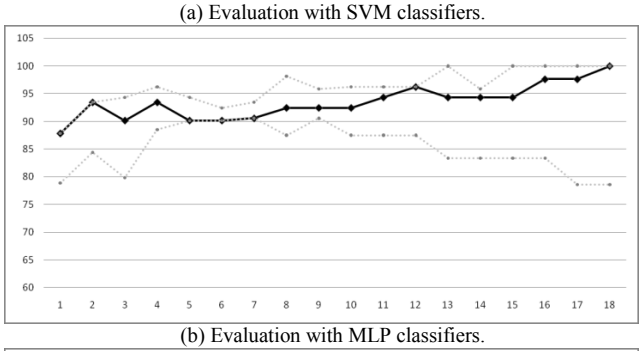
(e) Evaluation with DT classifiers.

Consensus Ranking Information Gain Chi-Square

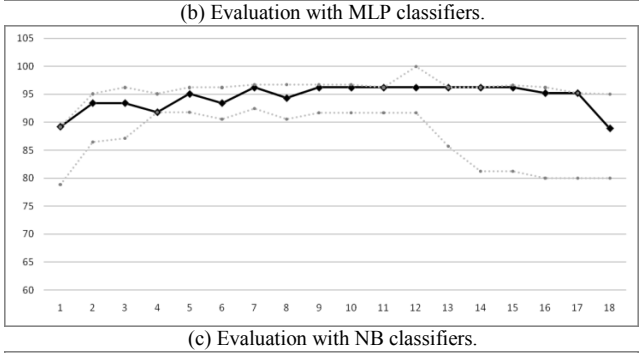
Figure 2. PI of the ranking methods.



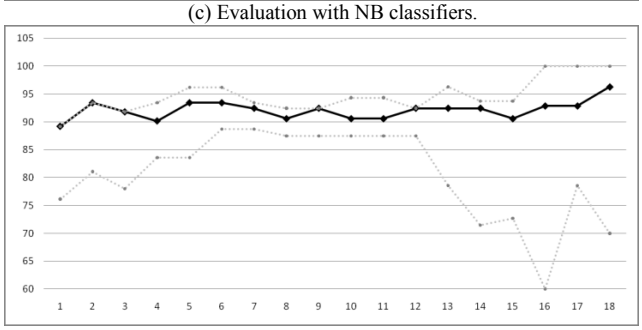
(a) Evaluation with SVM classifiers.



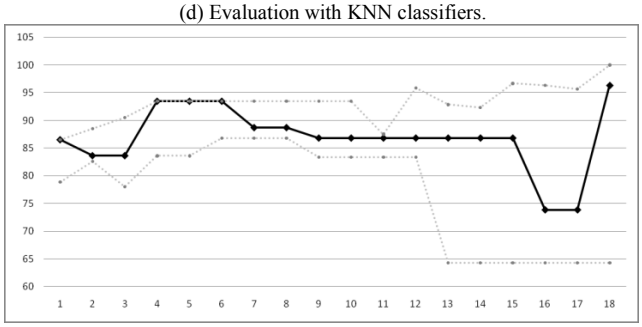
(b) Evaluation with MLP classifiers.



(c) Evaluation with NB classifiers.



(d) Evaluation with KNN classifiers.



(e) Evaluation with DT classifiers.

Consensus Ranking MIN MAX

Figure 3. Classification accuracy of the consensus ranking method.