IETE Technical Review
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/titr20

A Review of Ensemble Learning Based Feature Selection

Donghai Guan[a], Weiwei Yuan[a], Young-Koo Lee[a], Kamran Najeebullah[a] & Mostofa Kamal Rasel[a]
[a] Department of Computer Engineering, Kyung Hee University, Seoul, Korea
Published online: 23 Jun 2014.

To cite this article: Donghai Guan, Weiwei Yuan, Young-Koo Lee, Kamran Najeebullah & Mostofa Kamal Rasel (2014) A Review of Ensemble Learning Based Feature Selection, IETE Technical Review, 31:3, 190-198

To link to this article: http://dx.doi.org/10.1080/02564602.2014.906859

A Review of Ensemble Learning Based Feature Selection

Donghai Guan, Weiwei Yuan, Young-Koo Lee, Kamran Najeebullah and Mostofa Kamal Rasel

Department of Computer Engineering, Kyung Hee University, Seoul, Korea

ABSTRACT

Feature selection is an important topic in machine learning. In recent years, via integrating ensemble learning, the ensemble learning based feature selection approach has been proposed and studied. The general idea is to generate multiple diverse feature selectors and combine their outputs. This approach is superior to conventional feature selection methods in many aspects. Among them, its most prominent advantage is the ability to handle stability issue that is usually poor in existing feature selection methods. This review covers different issues related to ensemble learning based feature selection, which include the main modules, the stability measurement, etc. To the best of our knowledge, this is the first review that focuses on ensemble feature selection. It can be a useful reference in the literature of feature selection.

Keywords:

Feature selection, Ensemble learning, Stability.

1. INTRODUCTION

Machine learning techniques are widely used to analyse different types of data [1−5]. However, with the proliferation of large-scale data-sets, many machine learning techniques suffer from intractability problems. Although diverse factors cause intractability, the number of attributes, or dimensionality, is one of the major factors. Data dimensionality affects both the training and run-time phases of a learning system. Meanwhile, high dimensionality may cause dimensionality problems.

Feature selection is the technique to solve the dimensionality problem [6−13]. It retains only useful features via removing irrelevant and redundant features. It produces multiple benefits including improving classification performance, speeding up the data mining algorithm, and good insight of the problem through interpretation of the most relevant features.

The various feature selection algorithms differ mainly from two aspects: searching strategy and evaluation criterion. Based on these aspects, they can be divided into three categories: filter, wrapper, and embedded models. Although with diversity, in most existing feature selection methods, usually only a single feature selection process is involved.

In recent years, a new type of feature selection, called ensemble learning based feature selection, has been proposed and studied [14−35]. This approach integrates feature selection and ensemble learning. As shown in Figure 1, to integrate feature selection and ensemble learning, there exist two ways. One is to use ensemble learning for feature selection (FS_{en}). The other one is to utilize feature selection for ensemble learning (EN_{fs}) [36−38]. These two approaches look similar, but their goals are completely different. EN_{fs} is to use different feature subsets to construct an ensemble of diverse-based classifiers. The main research issue is to incorporate diversity into the objective function that supervises the search for the best features. But FS_{en} is designed to generate a good feature subset by combining multiple feature subsets based on the idea of ensemble learning.

This review focuses on the left branch of Figure 1, FS_{en}. The idea of FS_{en} is inspired due to the good performance of ensemble learning for supervised learning. In supervised learning, ensemble learning will generate multiple classifiers and then aggregate their classification results. It shows that this aggregated result is usually more accurate than the results from each individual classifier.

Similarly, the general idea of FS_{en} is to repeat the feature selection process for several times to generate diverse feature selectors, and aggregate their outputs. FS_{en} is expected to utilize the idea of ensemble learning to overcome the potential local optima problem of conventional single learning based feature selection. Various studies have proved that via integrating ensemble learning, FS_{en} is superior to conventional feature

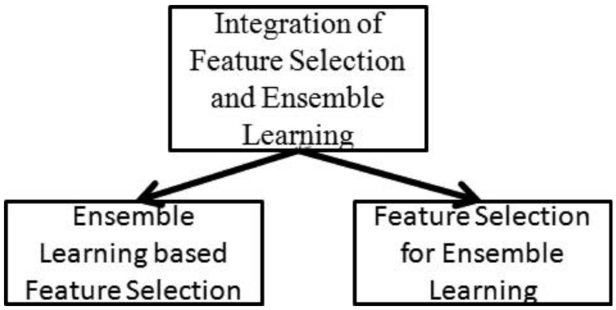

Figure 1: Two ways to integrate feature selection with ensemble learning.

selection from many aspects. Especially, it can improve the "stability" of feature selection, which usually cannot be satisfied by conventional feature selection.

This paper reviews the existing FS_{en} methods from different aspects, including the ways to generate diverse feature selectors, the ways to aggregate these selectors, and the ways to define the stability measurement.

Note that the reviews of feature selection already exist (e.g. [39–42]). This work differs from the existing review papers in the following aspects:

1. Existing reviews are aimed to cover all the different algorithms in feature selection. But because the idea of ensemble learning based feature selection is mainly studied in recent years, they are not covered by the most existing reviews.
2. To the best of our knowledge, this is the first review of ensemble learning based feature selection approaches.

The remaining of this paper is organized as follows. Section 2 introduces the concept and motivation of ensemble learning based feature selection. Subsequently, the various methods are presented in Section 3. In Section 4, we discuss some other topics that are related to FS_{en}. We finally conclude this review in Section 5.

2. ENSEMBLE LEARNING BASED FEATURE SELECTION

Feature selection chooses a subset of features among original feature spaces according to strategies that remove unrelated and redundant features. When fewer dimensions are obtained, the exploration of results is facilitated. Therefore, feature selection is widely used in many different data analysis tasks.

There are three kinds of feature selection methods: wrapper, filter, and embedded methods. Wrappers [43] choose features with high classification accuracies, as estimated by specified learning algorithms. Unlike wrappers, filters [44] evaluate the goodness of the proposed feature subset looking only at the intrinsic characteristics of the data, based on the relation of each single feature with the class label by the calculation of simple statistics computed from the empirical distribution. By taking classification accuracy into account, wrappers usually achieve better results than filters. Meanwhile, wrapper methods require more searching time, and are inefficient for large-scale problems. Finally, embedded methods [45] use internal information of the classification model to perform feature selection. Thus, they often provide a good trade-off between performance and computational cost.

The functions and usages of feature selection are diverse [46–47]. Because it is generally used as a preprocessing step for classification, classification accuracy is usually treated as the main goal of feature selection. For example, wrapper methods directly use classification accuracy to evaluate the goodness of features.

In addition to high accuracy, in recent years, the stability of feature selection has been recognized as another important issue [14–18, 20, 23, 27–28]. Stability is the measure to reflect the insensitivity of the feature selection algorithm to the variations of the training set. This issue is particularly important when using feature selection for knowledge discovery. Let us take the example of microarray analysis, wherein there are a big number of features. Feature selection is used to retain only important features and remove others. Finally, these selected features will be used by the biologists to explain the behaviours of microarray samples. Suppose we conduct feature selection for several times, if the training data are with variations, then the feature selection results with each time are also different. In terms of classification performance, all of the different feature subsets might be good. But such instability will result in the confidence drops to the domain experts. It is therefore necessary that the selected features are common to variations of training samples. Moreover, it also shows that usually the good stability can lead to the good classification/generalization accuracy.

FS_{en} is mainly proposed to improve the stability of feature selection. The role of ensemble learning in FS_{en} is similar with that for supervised learning [48]. For supervised learning, multiple predictors are generated and their predictions are fused to make the final decision. In FS_{en}, based on the training data, feature selection processes will run for several times to generate several diverse feature selectors. Later, these diverse feature selectors are aggregated into the final feature subset. It has been proved that FS_{en} is extremely important for the small sample size domains. For such domains, stability is more serious as usually several different feature subsets may yield equally optimal results, and ensemble feature selection may reduce the risk of choosing an unstable subset.

Furthermore, feature selection can be regarded as a searching process of the best features from the complete searching space. A single feature selection algorithm tends to select the local optima, while the ensemble feature selection has more chance to find the better features through aggregation of multiple feature selection processes. Finally, compared to single feature selection, the searching space of ensemble feature selection is enlarged so that the representation ability is also improved.

3. COMPONENTS IN ENSEMBLE LEARNING BASED FEATURE SELECTION

There are two essential steps in creating a feature selection ensemble. The first step is to generate a set of diverse feature selectors. Then, the second step is to aggregate the results of these feature selectors.

In this section, the two steps in Figure 2 will be introduced in detail. Moreover, as the main goal of ensemble feature selection is to improve the stability of feature selection, different stability measures will be discussed in this section. In essence, the two steps in Figure 2 and the stability measurements are dependent on each other. For example, the aggregation function can influence the stability of feature selection. Meanwhile, the stability measure can also limit the choices of aggregation functions.

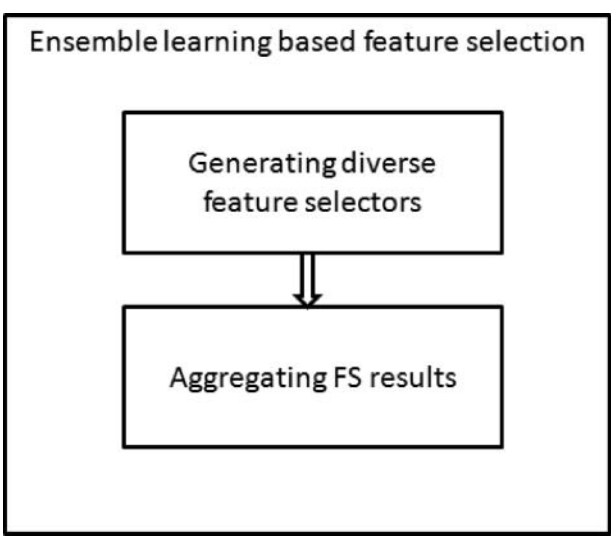

Figure 2: Components in ensemble learning based feature selection.

Table 1: The general scheme of ensemble feature selection based on data variation

Input: E (training set), A (feature selection algorithm)
Parameters: n (number of feature selectors)
Output: F2S (features to select)

(1) Form n different training sets ($E_1, E_2, \ldots, E_n$) based on E
(2) For $i = 1, \ldots, n$ do
(3) $FS_i = A(E_i)$
(4) End for
(5) F2S = Aggregation ($FS_1, FS_2, \ldots, FS_n$)

3.1 Generating Diverse Feature Selectors

Assume that n feature selectors are used in FS_{en}. Each feature selector $fs^i = \text{FS Function}^i(\text{Training Data}^i)$. Here FSFunction^i and TrainingData^i are the feature selection function and training data used for the ith feature selector.

In ensemble learning based feature selection, it is important to have enough diversity among different feature selectors. This is because if the feature selectors are very similar or identical, their combination will be similar to each feature selector and nothing can be achieved through their combination. As shown in the above equation, to make fs^i different, there are three approaches: making FS function different, making training data different, or making both of them different. Accordingly, these three methods are called data variation, function variation, and hybrid variation, respectively.

3.1.1 Data Variation Methods

The general scheme of data variation based FS_{en} is shown in Table 1.

The works in this category differ in the way to form different training sets (step (1) of Table 1). In the literature of ensemble learning, the popular methods include using different sample subsets (e.g. bagging [49], boosting [50]) or different feature subspaces (e.g. random subspace [51]). In the literature of FS_{en}, bagging is mainly used. Its algorithm is shown in Table 2.

Table 2: The bagging algorithm

Input: E (training set)
Parameters: n (number of feature selectors), m (number of training instances)
Output: The generated training subsets ($E_1, E_2, \ldots, E_n$)

(1) Set $E_1 = E_2 = \ldots E_n = \varnothing$
(2) For $i = 1, \ldots, n$ do
(3) For $j = 1, \ldots, m$
(4) Index $= m^*$ rand()
(5) $E_i = E_i \cup E^{\text{index}}$
(6) End for
(7) End for

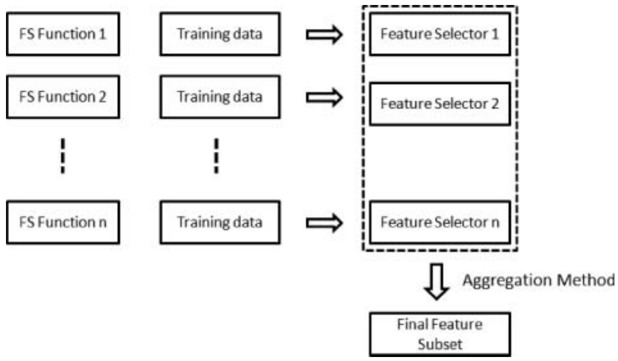

Figure 3: Function aggregation methods.

3.1.2 Function Variation Methods

For function variation methods, as shown in Figure 3, the training data are identical for each feature selector. The diversity of feature selectors is due to the diversity of feature selection functions.

Table 3 shows that for function variation methods, usually filters are employed. Recall that in addition to filter, wrapper is the other popular feature selection approach. Compared to filters, wrappers need much more computational costs. In function variation, as the feature selection process will run for several times. Therefore, filter is more suitable than wrapper.

The filter methods for function variation mainly include document frequency thresholding, chi-squared, information gain, ReliefF, mutual information, Kolmogorov−Smirnov statistic, deviance, geometric mean, area under the receiver operator characteristic (ROC) curve, area under the precision recall curve, signal to-noise (S2N), etc.

3.1.3 Hybrid Variation Methods

Hybrid variation methods believe that the diversity of individual data variation or function variation is not enough. Therefore, they combine these two variation methods. The recent studies of hybrid variation are shown in Table 4.

Dittman et al. [27] shows that the similarity between function variation and hybrid variation is much higher than their similarity between data diversity. In addition to have the highest similarity, function and hybrid variation methods show greater classification performance than data variation.

Because the size of training data is small, function and hybrid variation show outstanding performances. With small size, the diversity via data manipulation is very limited. Subsequently, the performance of data diversity is not satisfied. On the other hand, due to the relative large number of features, the function and hybrid variation can make more diversity for each feature selectors. This suggests that the size of training data is an important factor to choose the suitable method to generate multiple feature selectors.

3.2 Aggregation Functions

Considering there are n feature selectors in the FS_{en}. $FS_{en} = \{FS_1, FS_2, \ldots, FS_n\}$. According to the outcome in FS_i, feature selection algorithms can be divided into the following types:

1. Feature weighting algorithms. Each FS_i provides the weight for each feature. Therefore, each feature selection result, FS_i ($i = 1, 2, \ldots, n$), is a feature weighting vector. For feature weighting algorithms, the most common method is average the weights of each feature selector for a certain feature. Let FS_i^j represent the weight values of jth feature assigned by the ith feature selector. Then, the aggregated weight of jth feature is $FS^j = \frac{1}{n}\sum_{i=1:n} FS_i^j$.
2. Feature ranking algorithms. Each FS_i provides the rank for the features. Therefore, each feature

Table 3: The representative works that use function variation methods

Reference	Function variation	Domain	Aggregation method
Olsson et al. [35]	3 filter-based methods	Text data	Highest rank, Lowest rank, Average rank
Wang et al. [33]	18 filter-based methods	Software data-sets	Arithmetic mean
Wang et al. [52]	6 filter-based methods	Software data-sets	Arithmetic mean
Dittman et al. [27]	10 filter-based methods	Bioinformatics	Arithmetic mean

Table 4: The representative works that use hybrid variation methods

Reference	Function variation	Data variation	Aggregation method
Awada et al. [30]	10 filter-based methods	Bagging	Arithmetic mean
Dittman et al. [27]	10 filter-based methods	Bagging	Arithmetic mean

selection result, FS_i ($i = 1, 2, \ldots, n$) is a feature ranking vector. For feature ranking algorithms, the most common method is average the ranks of each feature selector for a certain feature. Let FS_i^j represent the rank values of jth feature assigned by the ith feature selector. Then, the aggregated rank of jth feature is $\mathrm{FS}^j = \frac{1}{n}\sum_{i=1:n}\mathrm{FS}_i^j$. Finally, the features can be ranked based on the aggregate value of each feature.

3. Feature subset selection algorithms. Each FS_i ($i = 1, 2, \ldots, n$) can be regarded as a vector of boolean values. If the jth feature is selected by FS_i, then the value of FS_i^j equals to 1. Otherwise, this value equals to 0. The aggregated weight of jth feature can still use the average aggregation method. Finally, the features can be ranked based on the aggregate weight of each feature.

In essence, the methods mentioned here can be generalized to the linear aggregation. In linear aggregation, the final rank of feature j is $\mathrm{FS}^j = \sum_{i=1:n} w(\mathrm{FS}_i^j)$. Here $w(\cdot)$ denotes a weighting function. If a linear aggregation is performed using $w(\mathrm{FS}_i^j) = \frac{\mathrm{FS}_i^j}{n}$, this means the arithmetic average of each ranking. By modifying $w(\mathrm{FS}_i^j)$, more complex aggregation methods can be proposed. For example, we can use a function such that more weight can be assigned to the feature selectors which have more reliability.

In addition to linear aggregation, other aggregation methods also can be used. Most of these methods are heuristic and show good performance in some applications. For example, in [35], the highest rank combination has been proposed for the feature ranking algorithm. In the highest rank combination, each feature's combined score is the highest rank achieved in any of the lists. This approach is suitable if different feature selectors place different sets of informative features near the top of their lists. This method can discount the negative information provided by the input. Similarly, the lowest and average rank can also be used [35].

3.3 Stability Measurements

Stability is the major goal of ensemble learning based feature selection. To evaluate it, subsampling-based methods have been proposed. The key idea of these methods is to see how the feature selection results will change due to the variation of training data. Suppose the training data-set $E = \{e_1, \ldots, e_m\}$ consists of m instances and N features. Then, a partial subset can be randomly drawn from data-set E. We define x as the ratio of data to choose from E. The subsampling process will be repeated for k times and thus k data-sets will be generated. Subsequently, feature selection is performed on each of the k data-sets, and a measure of stability is calculated. The calculation of stability is measured by comparing the outputs of the k feature selectors. The highest stability is achieved if the k feature selection results are completely identical. The total stability is the average over all pairwise similarity comparisons between the different feature selectors:

$$S_{\mathrm{tot}} = \frac{2\sum_{i=1}^{k}\sum_{j=i+1}^{k} S(\mathrm{FS}_i, \mathrm{FS}_j)}{k(k-1)}$$

where FS_i represents the feature selection result based on subsample i ($1 \le i \le k$), and $S(\mathrm{FS}_i, \mathrm{FS}_j)$ represents a similarity measure between FS_i and FS_j. Obviously, the total similarity depends on how to define the similarity measure. In the literature, the similarity measure is defined according to the outcomes of FS that include feature weighting, feature ranking, and feature subset selection. Note that the similarity measure can only be applied to two feature selection results with the same type. Let us define the number of features as N. Then, FS_i is a vector whose length is N. FS_i^j denotes (1) the weight of jth feature for feature weighting, (2) the rank of jth feature for feature ranking (1 for the worst rank and N for the best rank), and (3) 1 if the jth feature is selected (0 otherwise) for feature subset selection. The widely used definition of similarity measure is presented as follows:.

1. Feature weighting algorithms. Pearson correlation coefficient can be used as

$$S(\mathrm{FS}_i, \mathrm{FS}_j) = \frac{\sum_l (\mathrm{FS}_i^l - \mu_{f_i})(\mathrm{FS}_j^l - \mu_{f_j})}{\sqrt{\sum_l (\mathrm{FS}_i^l - \mu_{f_i})^2 \sum_l (\mathrm{FS}_j^l - \mu_{f_j})^2}}$$

2. Feature ranking algorithms. The Spearman rank correlation coefficient can be used as

$$S(\mathrm{FS}_i, \mathrm{FS}_j) = 1 - 6\sum_l \frac{(\mathrm{FS}_i^l - \mathrm{FS}_j^l)^2}{N(N^2-1)}$$

3. Feature subset selection algorithms. The Jaccard index can be used as

$$S(\mathrm{FS}_i, \mathrm{FS}_j) = \frac{|\mathrm{FS}_i \cap \mathrm{FS}_j|}{|\mathrm{FS}_i \cup \mathrm{FS}_j|} = \frac{\sum_l I(\mathrm{FS}_i^l = \mathrm{FS}_j^l = 1)}{\sum_l I(\mathrm{FS}_i^l + \mathrm{FS}_j^l > 0)}$$

In addition to the above stability measures, a number of other stability measurements have also been used.

Dunne et al. [53] present an approach based on Hamming distance:

$$H(\mathrm{FS}_i, \mathrm{FS}_j) = \sum_{x=1}^{l} |\mathrm{FS}_i^l - \mathrm{FS}_j^l|$$

Given a set k of feature selectors, the total Hamming distance (H_{tot}) is computed as follows:

$$H_{\mathrm{tot}} = \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} H(\mathrm{FS}_i, \mathrm{FS}_j)$$

Kuncheva [54] developed the idea of consistency index to measure similarity for feature subset selection algorithms. In this method, the consistency index is obtained as follows:

$$I_C(\mathrm{FS}_i, \mathrm{FS}_j) = \frac{dp - t^2}{t(p-t)}$$

where d is the cardinality of the intersection between FS_i and FS_j. The length of FS_i is t. This consistency index is between -1 and $+1$. If the two feature selectors are more similar, the consistency index will become greater.

4. DISCUSSIONS

This paper reviews the existing work of ensemble learning based feature selection. According to the usage of class labels of training data, feature selection can be divided into supervised and unsupervised. Supervised feature selection usually aims to maximize the predictive accuracy though selecting the relevant subset of features.

But in unsupervised learning, as the class labels are unknown, it becomes more challengeable to know which features should be selected. The general idea of unsupervised feature selection is to find relevant subsets of features that produce natural groupings by grouping similar objects together based on some similarity measure.

Although in the literature, the research of ensemble feature selection mainly focuses on supervised FS, the study on unsupervised feature selection also exists. Haytham et al. [19] proposed a method called random cluster ensemble, which estimates the out-of-bag feature importance from an ensemble of partitions. Each partition is constructed using a different bootstrap sample and a random subset of the features. The good performance of this method suggests that the study of unsupervised ensemble feature selection should get more attention.

Another topic to discuss is $\mathrm{EN}_{\mathrm{fs}}$. It has been proved that $\mathrm{EN}_{\mathrm{fs}}$ is an effective approach to generate an ensemble of accurate and diverse base classifiers. The general idea of $\mathrm{EN}_{\mathrm{fs}}$ is to vary the feature subsets and use them to generate diverse classifiers. For traditional feature selection, the goal is to find the best feature subset that needs considering both the learning task and the selected inductive learning algorithm. But for $\mathrm{EN}_{\mathrm{fs}}$, one additional goal is to find a set of feature subsets that can promote diversities among the base classifiers.

Similar to traditional feature selection, there are two factors when designing $\mathrm{EN}_{\mathrm{fs}}$: (1) search strategy and (2) fitness function. The search strategy searches for the feature subset which maximizes the output of fitness function. In traditional feature selection, the classification accuracy is mainly considered in the fitness function. But for $\mathrm{EN}_{\mathrm{fs}}$, the fitness function should reflect both the classification accuracy and the diversity. One of the fitness functions, proposed by Opitz [36], defines the fitness value as follows: $\mathrm{Fitness}_i = \mathrm{accu}_i + \alpha \cdot \mathrm{diversity}_i$,where α reflects the influence of diversity. $\mathrm{diversity}_i$ is the contribution of the ith classifier to the total ensemble diversity, which can be measured as the average pairwise diversity for all the pairs of classifiers including i.

$\mathrm{EN}_{\mathrm{fs}}$ and $\mathrm{FS}_{\mathrm{en}}$ are compared in Table 5. Through this table, the similarity and difference between them are summarized.

Table 5: Comparison between two ways of integrating feature selection and ensemble learning

		Feature selection for ensemble learning ($\mathrm{EN}_{\mathrm{fs}}$)	Ensemble learning for feature selection ($\mathrm{FS}_{\mathrm{en}}$)
Similarity		Integration of feature selection and ensemble learning	Integration of feature selection and ensemble learning
Difference	Goal	(1) Use feature selection to generate different feature subsets; these different feature subsets are used to construct an ensemble of diverse classifiers	(1) Generates a good feature subset by combining multiple feature subsets based on the idea of ensemble learning.
		(2) To achieve more accurate ensemble classifiers	(2) To obtain more stable feature subset
	Main research issues	(1) How to define search strategy	(1) How to generate diverse feature selectors
		(2) How to define fitness function	(2) How to aggregate these feature selectors

5. CONCLUSIONS

In this paper, the idea of ensemble learning based feature selection is reviewed. This is a relatively new idea and gets attention in recent years. Compared to traditional feature selection, ensemble feature selection has several advantages. It can overcome the potential local optima, enlarge the search space, and improve the generalization accuracy and stability. Although stability is usually neglected before, but in fact it is extremely important for knowledge discovery. Regarding to stability, ensemble feature selection has shown significant improvement.

Moreover, we also discuss the other two topics that are related to ensemble feature selection. One is ensemble feature selection for unsupervised learning. The other topic is feature selection for ensemble learning. We believe that the knowledge in these two topics can also contribute to the promotion of ensemble feature selection.

Funding

Dr. Young-Koo Lee is the corresponding author of this work. This work was supported by a grant from Kyung Hee University in 2010 [grant number KHU-20100168].

REFERENCES

1. I. Kononenko, "Machine learning for medical diagnosis: History, state of the art and perspective," *Artif. Intell. Med.*, Vol. 23, no. 1, pp. 89–109, Aug. 2001.
2. T. I. Tang, G. Zheng, Y. L. Huang, G. F. Shu, and P. T. Wang, "A comparative study of medical data classification methods based on decision tree and system reconstruction analysis," *Ind. Eng. Manag. Syst.*, Vol. 4, pp. 102–8, Jun. 2005.
3. W. Wajs, P. Wais, M. Swiecicki, and H. Wojtowicz, "Artificial immune system for medical data classification," in *Proceedings of the 2005 International Conference on Computational Science*, Atlanta, GA, May 22–25, 2005, pp. 810–12.
4. K. Saastamoinen, and J. Ketola, "Medical data classification using logical similarity based measures," in *Proceedings of the 2006 IEEE Conference on Cybernetics and Intelligent Systems*, Bangkok, Jun. 7–9, 2006, pp. 1–5.
5. K. Polat, S. Sahan, H. Kodaz, and S. Gunes, "Breast cancer and liver disorders classification using artificial immune recognition system (AIRS) with performance evaluation by fuzzy resource allocation mechanism," *Expert Syst. Appl.*, Vol. 32, no. 1, pp. 172–83, Jan. 2007.
6. T. H. Cheng, C. P. Wei, and V. S. Tseng, "Feature selection for medical data mining: comparisons of expert judgment and automatic approaches," in *Proceedings of the 19th IEEE International Symposium on Computer-Based Medical Systems*, pp. 165–70, 2006.
7. K. Polat, and S. Gunes, "A new feature selection method on classification of medical datasets: Kernel F-score feature selection," *Expert Syst. Appl.*, Vol. 36, no. 7, pp. 10367–73, Sep. 2009.
8. L. Chuang, H. Chang, C. Tu, and C. Yang, "Improved binary PSO for feature selection using gene expression data," *Comput. Biol. Chem.*, Vol. 32, no. 1, pp. 29–38, Feb. 2008.
9. Y. Huang, P. McCullagh, N. Black, and R. Harper, "Feature selection and classification model construction on type 2 diabetic patients' data," *Artif. Intell. Med.*, Vol. 41, no. 3, pp. 251–62, Nov. 2007.
10. R. Abdel-Aal, "Improved classification of medical data using abductive network committees trained on different feature subsets," *Comput. Methods Programs Biomed.*, Vol. 80, no. 2, pp. 141–53, Nov. 2005.
11. Y. Marinakis, G. Dounias, and J. Jantzen, "Pap smear diagnosis using a hybrid intelligent scheme focusing on genetic algorithm based feature selection and nearest neighbor classification," *Comput. Biol. Med.*, Vol. 39, no. 1, pp. 69–78, Jan. 2009.
12. S. N. Ghazavi, and T. W. Liao, "Medical data mining by fuzzy modeling with selected features," *Artif. Intell. Med.*, Vol. 43, no. 3, pp. 195–206, July 2008.
13. R. Abdel-Aal, "GMDH-based feature ranking and selection for improved classification of medical data," *J. Biomed. Inform.*, Vol. 38, no. 6, pp. 456–68, Dec. 2005.
14. Y. Saeys, T. Abeel, and Y. Peer, "Robust feature selection using ensemble feature selection techniques," in *Proceedings of the ECML PKDD*, Vol. 5212, 2008, pp. 313–25.
15. S. Loscalzo, L. Yu, and C. Di, "Consensus group stable feature selection," in *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, Paris, Jun. 28–July 1, 2009, pp. 567–75.
16. Y. Li, S. Y. Gao, and S. C. Chen, "Ensemble feature weighting based on local learning and diversity," in *Proceedings of the AAAI Conference on Artificial Intelligence*, Toronto, 2012, pp. 1019–25.
17. A. Woznica, .P Nguyen, and A. Kalousis, "Model mining for robust feature selection," in *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, Beijing, Aug. 12–16, 2012, pp. 913–21.
18. Y. Li, S. Huang, S. Chen, and J. Si, "Stable L2-Regularized ensemble feature weighting," *Lect. Notes Comput. Sci.*, Vol. 7872, pp. 167–78, 2013.
19. H. Elghazel, and A. Aussem, "Unsupervised feature selection with ensemble learning, *Mach. Learn.*, Apr. 2013, 10.1007/s10994-013-5337-8.
20. T. Abeel, T. Helleputte, Y. Peer, P. Dupont, and Y. Saeys, "Robust biomarker identification for cancer diagnosis with ensemble feature selection methods," *Bioinformatics*, Vol. 26, no. 3, pp. 392–98, Feb. 2010.
21. S. Landeghem, T. Abeel, Y. Saeys, and Y. Peer, "Discriminative and informative features for biomolecular text mining with ensemble feature selection," *Bioinformatics*, Vol. 26, no. 18, pp. 554–60, Sep. 2010.
22. V. BoloN-Canedo, N. SaNchez-MaronO, and A. Alonso-Betanzos, "An ensemble of filters and classifiers for microarray data classification," *Pattern Recogn.*, Vol. 45, no. 1, pp. 531–39, Jan. 2012 January.
23. W. Awada, T. M. Khoshgoftaar, D .Dittman, and R. Wald, "The Effect of Number of Iterations on Ensemble Gene Selection," in *Proceedings of the 11th International Conference on Machine Learning and Applications (ICMLA)*, Florida, Dec. 12–15, 2012, pp. 198–203.
24. P. Yang, Y. Hwa Yang, B. B. Zhou, and A. Y. Zomaya, "A review of ensemble methods in bioinformatics," *Curr. Bioinform.*, Vol. 5, no. 4, pp. 296–308, Dec. 2010.
25. H. Liu, L. Liu, and H. Zhang, "Ensemble gene selection by grouping for microarray data classification," *J. Biomed. Inform.*, Vol. 43, no. 1, pp. 81–7, 2010.
26. A. C. Haury, P. Gestraud, and J. P. Vert, "The influence of feature selection methods on accuracy, stability and interpretability of molecular signatures, 2011," *PLoS ONE*, Vol. 6, no. 12, pp. e28210. doi:10.1371/journal.pone.0028210.

27. J. Dittman, M. Khoshgoftaar, R. Wald, and A. Napolitano, "Comparing two new gene selection ensemble approaches with the commonly-used approach," in *Proceedings of the 11th International Conference on Machine Learning and Applications*, Florida, Dec. 12–15, 2012, pp. 184–91.
28. R. C. Prati, "Combining feature ranking algorithms through rank aggregation," in *Proceedings of the IEEE World Congress on Computational Intelligence*, Brisbane, Jun. 10–15, 2012, pp. 1–8.
29. R. Wald, T. M. Khoshgoftaar, D. Dittman, W. Awada, and A. Napolitano, "An extensive comparison of feature ranking aggregation techniques in bioinformatics", in *Proceedings of the IEEE 13th International Conference on Information Reuse and Integration*, Las Vegas, Aug. 8–10, 2012, pp. 377–84.
30. W. Awada, T. M. Khoshgoftaar, D. Dittman, R. Wald, and A. Napolitano, "A review of the stability of feature selection techniques for bioinformatics data" in *Proceedings of the 13th International Conference on Information Reuse and Integration*, Las Vegas, Aug. 8–10, 2012, pp. 356–63.
31. Z. He, and W. Yu, "Stable feature selection for biomarker discovery," *Comput. Biol. Chem.*, Vol. 34, pp. 215–25, Aug. 2010.
32. H. Liu, L Liu, and H. Zhang, "Ensemble gene selection for cancer classification," *Pattern Recogn.*, Vol. 43, pp. 2763–72, Feb. 2010.
33. H. Wang, T. M. Khoshgoftaar, and A. Napolitano, "A comparative study of ensemble feature selection techniques for software defect prediction," in *Proceedings of the 9th International Conference on Machine Learning and Applications (ICMLA)*, Washington, DC, Dec. 12–14, 2010, pp. 135–40.
34. M. Netzer, and G. Millonig, "A new ensemble-based algorithm for identifying breath gas marker candidates in liver disease using ion molecule reaction mass spectrometry," *Bioinformatics*, Vol. 25, no. 7, pp. 941–7, Mar. 2009.
35. J. O. S. Olsson, and D. W. Oard, "Combining feature selectors for text classification," in *Proceedings of the 15th ACM International Conference on Information and Knowledge Management*, Arlington, VA, Nov. 5–11, 2006, pp. 798–9.
36. D. Opitz, "Feature selection for ensembles," in *Proceedings of the 16th National Conference on Artificial Intelligence (AAAI)*, Orlando, FL, July 18–22, 1999, pp. 379–84.
37. A. Tsymbal, S. Puuronen, and D Patterson, "Ensemble feature selection with the simple Bayesian classification information fusion," *Information Fusion*, Vol. 4, no. 2, pp. 87–100, Jun. 2003.
38. A. Tsymbal, M. Pechenizkiy, and P. Cunningham, "Diversity in search strategies for ensemble feature selection," *Inform. Fusion*, Vol. 6, no. 1, pp. 83–98, Mar. 2005.
39. Z. Zhao, "Spectral feature selection for mining ultrahigh dimensional data," Ph.D. thesis, Arizona State Univ., Tempe, AZ, 2010.
40. H. Liu, and L. Yu, "Toward integrating feature selection algorithms for classification and clustering", *IEEE Trans. Knowl. Data Eng.*, Vol. 17, pp. 494–502, Apr. 2005.
41. I. Guyon, S. Gunn, M. Nikravesh, and L. Zadeh, *Feature extraction, foundations and applications*. New York: Springer, Physica-Verlag; 2006.
42. I. Guyon, and A Elisseeff, "An introduction to variable and feature selection," *J Mach Learn. Res.*, Vol. 31, pp. 1157–82, Mar. 2003.
43. R. Kohavi, and G. John, "Wrappers for feature selection," *Artif. Intell.*, Vol. 97, no. 1, pp. 273–324, Mar. 1997.
44. N. Marono, A. Betanzos, and M. Sanroman, "Filter methods for feature selection, a comparative study," *Lect. Notes Comput. Sci.*, Vol. 4881, pp 178–87, 2007.
45. R. Uriarte, and S. Andres, "Gene selection and classification of microarray data using random forest," *BMC Bioinform.*, Vol. 7, pp. 3, Jan. 2006.
46. I. Inza, P. Larranaga, R. Blanco, and A. J. Cerrolaza, "Filter versus wrapper gene selection approaches in DNA microarray domains," *Artif. Intell. Med.*, Vol. 31, pp. 91–103, Jun. 2004.
47. G. Forman, "An extensive empirical study of feature selection metrics for text classification," *J. Mach. Learn. Res.*, Vol. 3, pp. 1289–1305, 2003 March.
48. T. Dietterich, "Ensemble methods in machine learning," *in Proc. 1st Int. Workshop Mult. Classifier Syst.*, Cagliari, Italy, Jun. 21–23, 2000, pp. 1–15.
49. L. Breiman, "Bagging predictors," *Mach. Learn.*, Vol. 24, no. 2, pp. 123–40, 1996.
50. Y. Freund, and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *J. Comput. Syst. Sci.*, Vol. 55, no. 1, pp. 119–39, Aug. 1997.
51. T. K. Ho, "The random subspace method for constructing decision forests," *IEEE Trans. Pattern Anal. Mach. Intell.*, Vol. 20, no. 8, pp. 832–44, Aug. 1998.
52. H. Wang, T. M. Khoshgoftaar, and K. Gao, "Ensemble feature selection technique for software quality classification," in *Proceedings of the 22nd International Conference on Software Engineering and Knowledge Engineering*, 2010, pp. 215–20.
53. K. Dunne, P. Cunningham, and F. Azuaje, "*Solutions to instability problems with sequential wrapper-based approaches to feature selection*," Dept. Comput. Sci., Trinity Coll., Dublin, Tech. Rep. TCD-CS-2002-28, 2002.
54. L I. Kuncheva, "A stability index for feature selection," in *Proceedings of the 25th Conference on Artificial Intelligence Applications*, Innsbruck, Feb. 12–14, 2007, pp. 390–95.

Authors

Donghai Guan received his PhD degree in computer science from Kyung Hee University (KHU), South Korea, in 2009. From 2009, he was a postdoc and research professor in KHU. Since February 2011, he has been an assistant professor in Harbin Engineering University, China. Since March 2012, he has been an assistant professor in Kyung Hee University, Korea. His research interests are machine learning and data mining.

E-mail: donghai@oslab.khu.ac.kr

Weiwei Yuan received her PhD degree in computer science from Kyung Hee University, South Korea, in 2010. Since September 2010, she has been an assistant professor in Harbin Engineering University, China. Since September 2012, she has been a postdoctoral fellow at Computer Engineering Department, Kyung Hee University. Her research interests are social computing, information security, and machine learning.

E-mail: yuanweiwei00@khu.ac.kr

Young-Koo Lee received his BSc, MSc and PhD in Computer Science from Korea Advanced Institute of Science and Technology, Korea in 2002. He was Postdoctoral Research Associate at Deptartment of Computer Science, University of Illinois at Urbana-Champaign, Illinois, US, from September 2002 to February 2004. He has been a Professor in the Department of Computer Engineering at Kyung Hee University, Korea since 2004. His research interests include ubiquitous data management, data mining, and databases. He is a member of the IEEE, the IEEE Computer Society, and the ACM.

E-mail: yklee@khu.ac.kr

Kamran Najeebullah received his BSc degree from Gomal University, Pakistan in 2008. He has been a Master's candidate at the Department of Computer Engineering, Kyung Hee University, South Korea since September 2012. His research interests include big graph processing, data mining, web mining and graph databases

E-mail: kamran@khu.ac.kr

Mostofa Kamal Rasel received his BSc degree from Islamic University of Technology, Bangladesh in 2009. He has been a Master's candidate at the Department of Computer Engineering, Kyung Hee University, South Korea since September 2012. His research interests include massive graph mining, data mining, and query processing techniques on modern storage drive.

E-mail: rasel@khu.ac.kr

DOI: 10.1080/02564602.2014.906859;