J Med Syst (2012) 36:2841–2847
DOI 10.1007/s10916-011-9762-6

ORIGINAL PAPER

Design Ensemble Machine Learning Model for Breast Cancer Diagnosis

Sheau-Ling Hsieh · Sung-Huai Hsieh · Po-Hsun Cheng · Chi-Huang Chen · Kai-Ping Hsu · I-Shun Lee · Zhenyu Wang · Feipei Lai

Received: 22 May 2011 / Accepted: 19 July 2011 / Published online: 3 August 2011

Abstract In this paper, we classify the breast cancer of medical diagnostic data. Information gain has been adapted for feature selections. Neural fuzzy (NF), k-nearest neighbor (KNN), quadratic classifier (QC), each single model scheme as well as their associated, ensemble ones have been developed for classifications. In addition, a combined ensemble model with these three schemes has been constructed for further validations. The experimental results indicate that the ensemble learning performs better than individual single ones. Moreover, the combined ensemble model illustrates the highest accuracy of classifications for the breast cancer among all models.

Keywords Ensemble learning · Neural fuzzy · KNN · Quadratic classifier · Information gain

Introduction

The diagnoses of diseases depend upon tests performed on patients. The actual causes of the illness can be difficult to identify or obtain even after undertaking a series of tests. Under the circumstances, enhanced approaches can assist physicians while classifying, figuring out the senses of the collected data.

Supervised machine learning approaches are promising for analyzing diagnostic medical data as well as microarray gene expression data. The learning abilities to construct classifiers or hypotheses can explain complex relationships among the data. Moreover, ensemble learning [1] is a technology to combine individual classifiers to classify the data. The technique can achieve higher accuracy of classifications [2].

S.-L. Hsieh · I.-S. Lee
Network and Computer Centre, National Chiao Tung University, Hsinchu, Taiwan

F. Lai
Department of Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan

F. Lai
Department of Electrical Engineering, National Taiwan University, Taipei, Taiwan

C.-H. Chen · K.-P. Hsu
Network and Computer Centre, National Taiwan University, Taipei, Taiwan

F. Lai
Graduate Institute of Biomedical Electronics and Bioinformatics, National Taiwan University, Taipei, Taiwan

Z. Wang
Computing Laboratory, Oxford University, Oxford, UK

P.-H. Cheng (✉)
Department of Software Engineering, National Kaohsiung Normal University, Kaohsiung, Taiwan
e-mail: cph@nknu.edu.tw

S.-H. Hsieh
Department of Computer Science and Information Engineering, Providence University, Taichung, Taiwan

In the study, we pursue Information Gain (IG) to fulfill obtaining the useful features. We propose four ensemble models to analyze the breast cancer diagnostic medical data. NF, KNN and QC single models, as well as their ensemble ones, i.e., NFE, KNNE, QCE, have been developed for classifications. In addition, according to the three classifiers, the fourth combined ensemble model is generated.

The paper is organized as followings. In Section "Methodology", we address the backgrounds, methodologies of constructing the three classifiers and their ensemble ones. The combined ensemble model is described in Section "The ensemble model contains NF, KNN, QC classifiers" especially. Experimental results are presented, illustrated clearly in Section "Experimental result" for the breast cancer. The comparison among the current research with previous ones is elaborated in the Section as well. Finally, the paper concludes in Section "Conclusion and future work".

Methodology

Information gain

Information Gain (IG) technique adapts the concept of Shannon entropy [3]. Given an entropy E to calculate the correlation in a training dataset, it also can estimate the usefulness of a feature while classifying the training data. The measurement is simply the expected reduction in entropy caused by partitioning the data based on the feature, i.e., Information Gain [4, 5].

The entropy of a variable X is defined as (1):

$$H(X) = -\sum_i P(x_i)\log_2(P(x_i)) \tag{1}$$

The entropy of X after observing values of another variable Y is defined as

$$H\left(\frac{X}{Y}\right) = -\sum_j P(y_j)\sum_i P\left(\frac{x_i}{y_j}\right)\log_2\left(P\left(\frac{x_i}{y_j}\right)\right) \tag{2}$$

where P(xi) is the prior probabilities for all values of X, and P(xi | yes) is the posterior probabilities of X given the values of Y . The amount by which the entropy of X decreases reflects additional information about X provided by Y and is called information gain, given by

$$IG\left(\frac{x}{y}\right) = H(X) - H\left(\frac{X}{Y}\right) \tag{3}$$

Given a set of medical expression data M, the information gain of a data i is

$$IG(M,i) = E(M) - \sum_{v\in V(i)} \frac{M_v}{M}E(M_v) \tag{4}$$

where V (i) is the set of all possible values of feature i, Mv is the subset of M for which feature i has value v, E(M) is the entropy of the entire set, and E(Mv) is the entropy of the subset Mv. The entropy function E is defined by:

$$E = \sum_{j=1}^{c} -\frac{|Cj|}{|C|}\log_2\frac{|Cj|}{|C|} \tag{5}$$

where |Cj | is the number of samples in class Cj, and the |C| is the sum of |Cj |. The entropy is supposed to provide the information required in bits, and is traditionally used to deal with boolean valued features (hot/cold, true/false, etc.). Fortunately, this method can be extended to handle the data with continuous valued features.

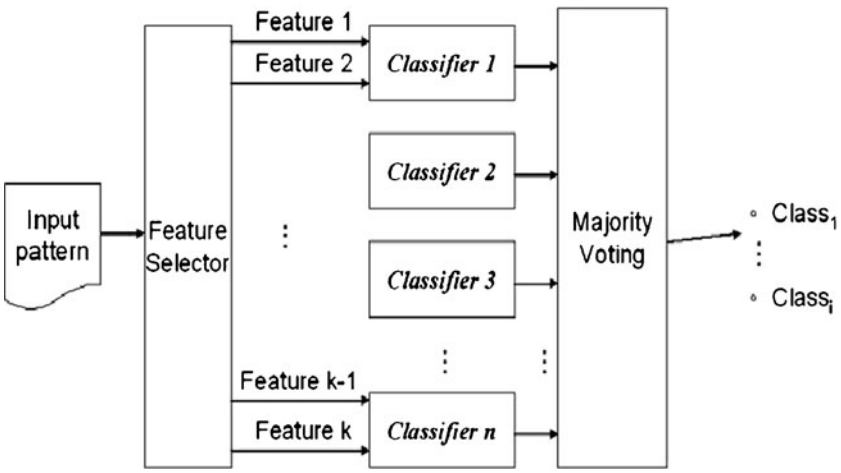

Fig. 1 Structure of an ensemble classifier model

Ensemble classifier model

There are increasing number of researchers attract their attentions in ensemble machine learning. Several ensemble techniques have been proposed. In addition, it has strong evidences that they can significantly enhance the accuracy of classifications. In here, we construct four different ensemble models by combining several single NF models, KNN models and QC models to learn the same data with different subset of medical data. The approach can allow the NF models or QC models to study more diagnostic medical data when a small subset of medical data cannot fully represent the whole. Moreover, the ensemble model anticipates a better classification performance.

In Fig. 1, the main structure of an ensemble classifier model is depicted. In the model, it contains N single classifiers; each single model has R inputs. Thus, the whole model can input R*N features. The output of the ensemble is calculated by simple majority voting (MV). The classifiers of ensemble model can be individual NF, KNN or QC model. Therefore, we establish three ensemble models: neural fuzzy ensemble (NFE) [6], k nearest neighbor ensemble (KNNE), as well as quadratic classifier ensemble (QCE) models for classifications.

The ensemble model contains NF, KNN, QC classifiers

The fourth ensemble model consists of clusters of classifiers; each cluster encompasses three different single classifiers: i.e., NF, KNN and QC models. The input features applied to individual single classifier are repeated and identical per cluster. The output strategy for the model is Majority Voting. The main structure of the ensemble model is shown in Fig. 2.

Table 1 The feature of breast cancer data set

Number #	Attribute	Domain
1	Sample code number	Id number
2	Clump Thickness	1–10
3	Uniformity of Cell Size	1–10
4	Uniformity of Cell Shape	1–10
5	Marginal Adhesion	1–10
6	Single Epithelial Cell Size	1–10
7	Bare Nuclei	1–10
8	Bland Chromatin	1–10
9	Normal Nucleoli	1–10
10	Mitoses	1–10
11	Class	2 for Benign, 4 for Malignant

Training and testing strategy

For classifications, normally the data are divided or allocated in three subsets, i.e., training, testing and validation data. However, for diagnostic medical expression data sets, it is difficult to determine which model has better accuracy than others with limited amount of samples. We eliminate the traditional training strategies of machine learning according to the three subsets approach. Any incorrect classification can induce the accuracy degradation dramatically. The variances of accuracy cannot represent the true classifications. Due to the fact that the number of samples is limited, each subset of them cannot fully represent the space. It is not appropriate to train on one space, but test on another very different one. Therefore, in order to train as many samples as possible [7], another strategy has been considered; leave one out cross validation (LOOCV). We divide all samples randomly into K distinct

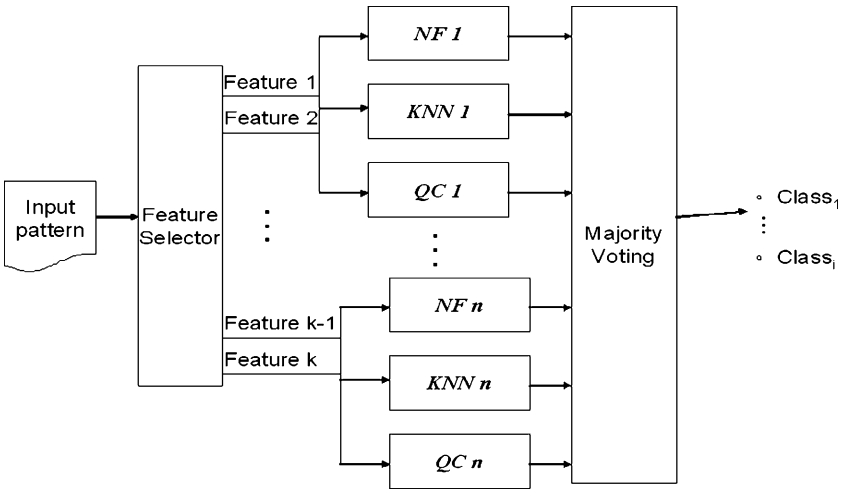

Fig. 2 Ensemble model contains NF, KNN and QC

Table 2 The ranked feature select by IG of breast cancer data set

Rank	Feature number	Feature
1	3	Uniformity of Cell Size
2	4	Uniformity of Cell Shape
3	7	Bare Nuclei
4	8	Bland Chromatin
5	6	ingle Epithelial Cell Size
6	9	Normal Nucleoli
7	2	Clump Thickness
8	5	Marginal Adhesion
9	10	Mitoses
10	1	Sample code number

subsets, where K equals to the number of samples. Train the model using K−1 subsets, and test the training performance on the Kth sample. The LOOCV accuracy is obtained by:

$$\text{LOOCV accuracy} = \frac{Acs}{K} \tag{6}$$

where Acs is the number of correctly classified samples in K experiments.

Experimental result

The breast cancer data set we used was retrieved from the UCI Machine Learning Repository [8, 9] in the University of Wisconsin Hospital database. This data set is medical diagnosis data and includes 699 samples. All samples can categorize into two subtypes: 458 samples of benign and 241 samples of malignant. The features are computed form a digitized image of a fine needle aspirate (FNA) of a breast mass. There are ten features in the data set and number 11 is the class of the data, as the Table 1 shows.

Feature selection results

We list the ranked features selected by Information Gain (IG) of breast cancer data set. In the breast cancer data set, we list all the rank of features selected by using IG method

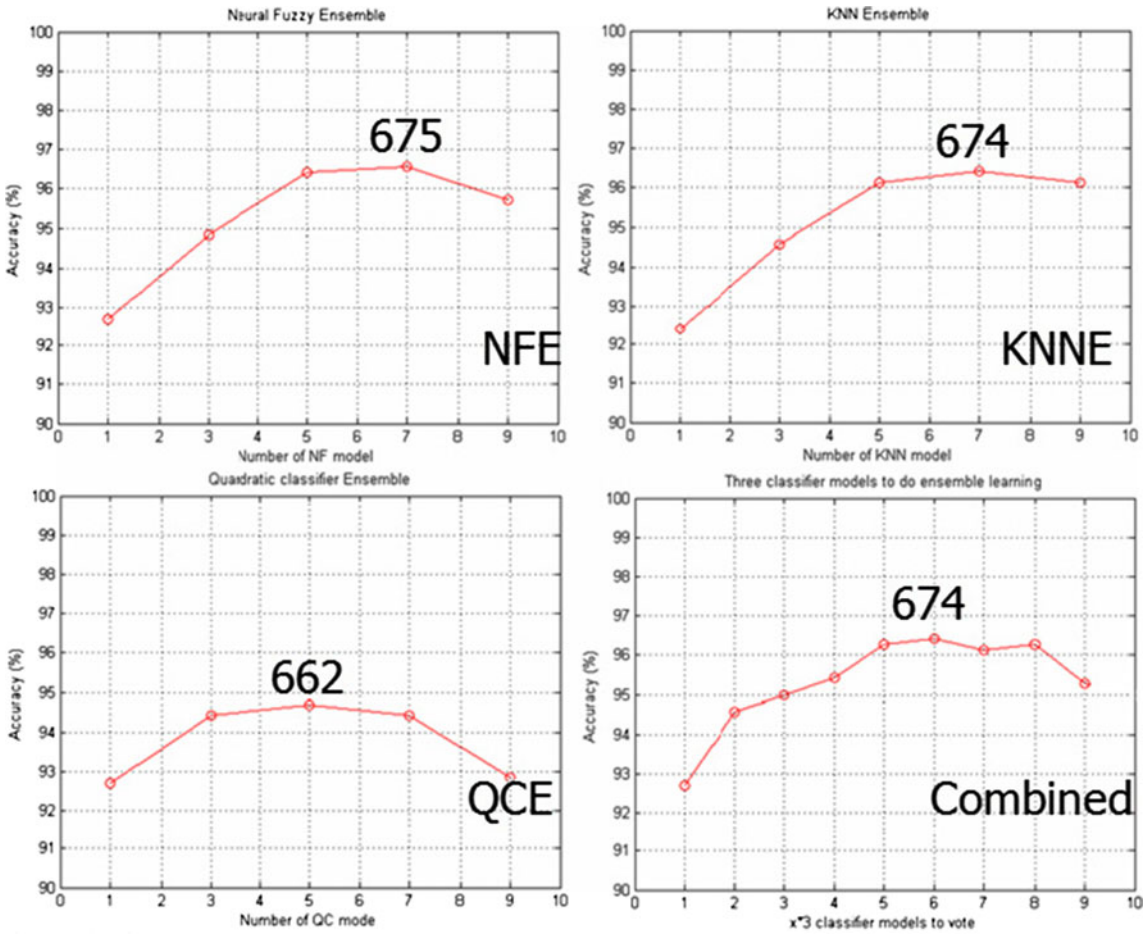

Fig. 3 The four ensemble models with 1 input feature

and were later for classification. Table 2 shows the top 10 features samples.

The number of selected features and LOOCV strategy are constituted to evaluate the accuracy as well as effectiveness of the models. In the diagrams, the vertical coordinate represents the correctness of classification. Each histogram indicates individual classifying having input features ranked in descending order. The dashed line above illustrates the classification correctness of the ensemble model. The final results are generated according to prior numbers of single classifiers voting (equally weighted). We compare experimental results of the four ensemble models, i.e., NFE, KNNE, QCE, and the combined ensemble model. The outcomes are described clearly in Figs. 3, 4 and 5 respectively.

In Fig. 3, the ensemble models classify on breast cancer data set, when the number of input feature for each classifier is 1. In the diagram, it indicates that the NFE model achieves the highest accuracy at 96.5% having 675 correct counts, when the input feature is 7.

Figure 4 shows the four ensemble models classify on breast cancer data set, when the number of input feature for each classifier is 2. In the diagram, it indicates that the combined ensemble model achieves the highest accuracy at 97.14% having 679 correct counts, when the input feature is 9.

Figure 5 depicts the four ensemble models classify on breast cancer data set, when the number of input feature for each classifier is 3. In the diagram, it indicates that the combined ensemble model achieves the highest accuracy at 96.71% having 676 correct counts, when the input feature is 9.

Comparisons

The classification accuracies comparing with other researches are listed clearly in followed table. The experimental data in the tables are retrieved from the same database repository [8]. In Table 3, apparently, it indicates the current research achieves the higher accuracy with the combined ensemble model containing NF, KNN and QC classifiers, at 97.14%. The highest accuracy is Seral Sahan's approach at 99.14%.

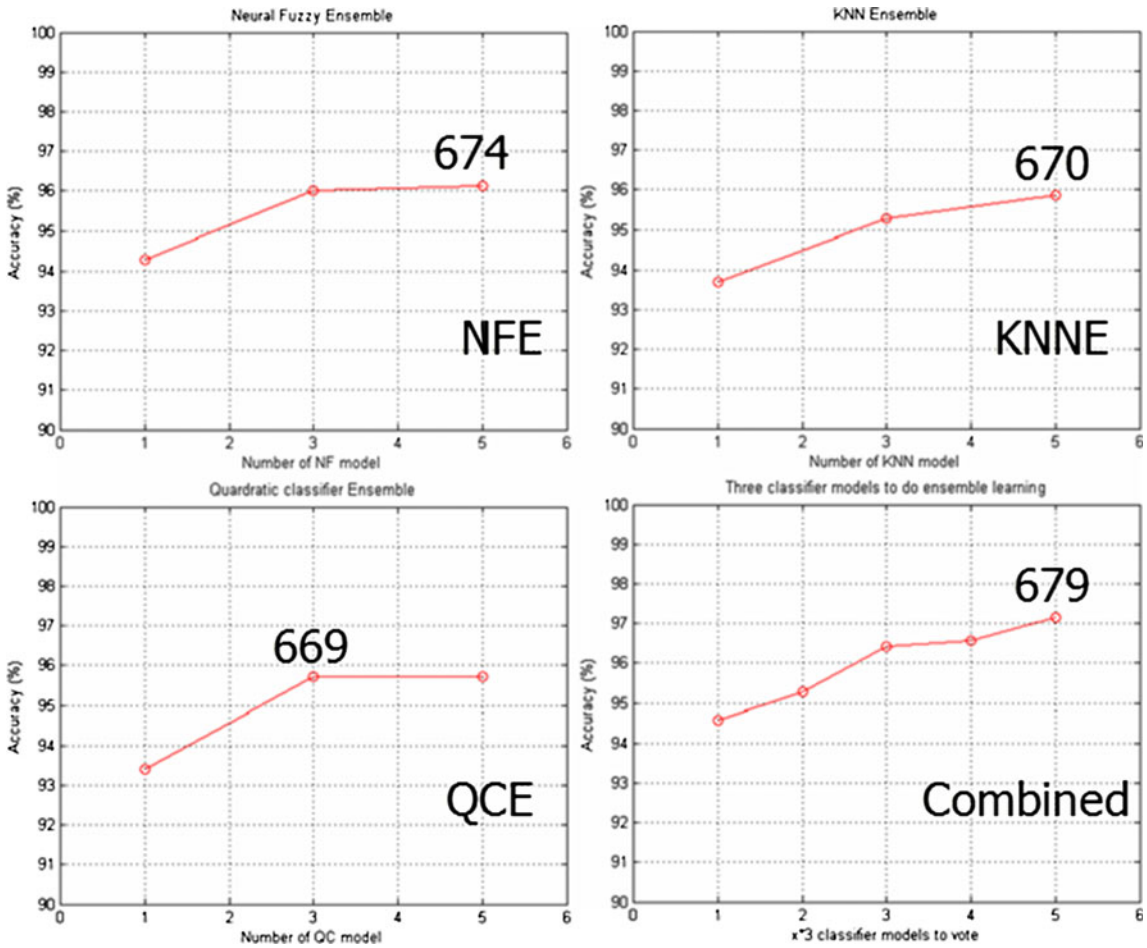

Fig. 4 The four ensemble models with 2 input features

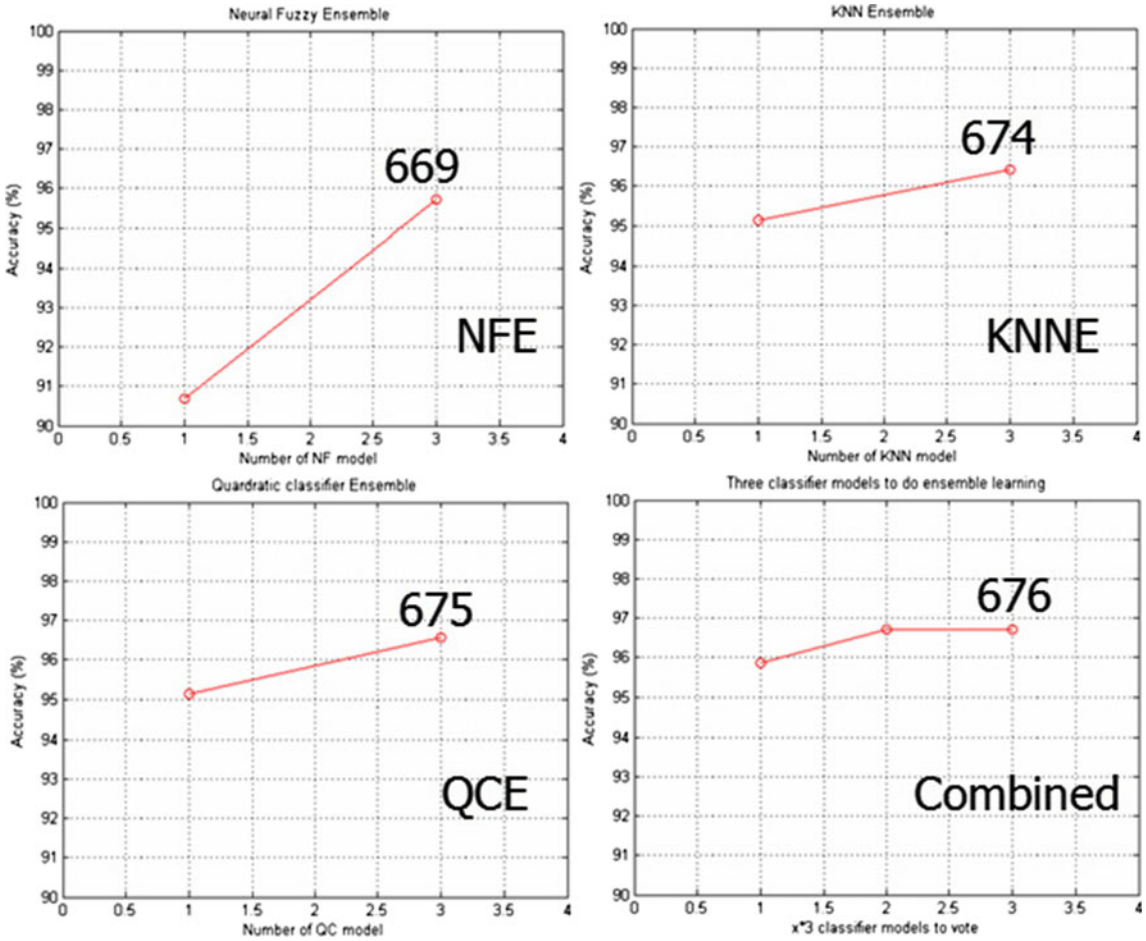

Fig. 5 The four ensemble models with 3 input features

Conclusion and future work

In the paper, we have constructed ensemble models for classifications. According to the study, we applied the NF, KNN and QC models for cancer data classifications. During the classification processes, it indicates that the NF and QC model cannot allow input a large number of features because of high computational costs. In order to target the problem, we adapt ensemble models, i.e., NFE, KNNE, QCE as well as combined model for classification. Therefore, the experimental results indicate NFE and QCE can cope with large data sets. And the ensemble classifier got better accuracy than individual classifier obtained on most tested data sets.

There also one efficient features selection method adopted in our experimental study, i.e., Information Gain

Table 3 The classification accuracy compare with other methods of breast cancer data set

Author (Year)	Method	Acc (%)	Author (Year)	Method	Acc (%)
Quinlan (1996) [10]	C4.5	94.74%	Our research (2010)	Single NF	94.28%
				Single KNN	96.42%
Amanda J.C. Sharkey (1998) [11]	NNE	96.50%		Single QC	94.50%
Abonyi and Szeifert (2003) [12]	Fuzzy clustering	95.57%		NFE	96.56%
				KNNE	96.42%
Gonzalo Martínez-Muñoz et al. (2005) [13]	Boosting C4.5	96.80%		QCE	96.57%
	Bagging C4.5	96.10%		Ensemble (NF KNN QC)	97.14%
Seral Sahan et al. (2007) [14]	Fuzzy-AIS-knn	99.14%			

(IG). LOOCV accuracy was used to evaluate the effectiveness of the classifiers, because traditional strategies could not well distinguish the ability of different classifiers by using such a small number of training patterns of medical expression data.

According to the experimental analyses, in order for an ensemble model to be effective and efficient, it should constitute diversified base classifiers as well as feature selections for classifications. Apparently, in the research, the combined ensemble models containing three different base classifiers demonstrate higher accuracies, on average, comparing with other homogeneous ensemble ones. However, the feature selection process can be sensitive to the diversity of data being classified. Currently, the input features are sequentially selected based on the Information Gain ranking with equally weighted voting mechanism. A combined approach to integrate diversity of base classifiers, the best feature selection algorithms, as well as dynamic voting techniques that is germane to both learning tasks and classifications is the future challenges. The tradeoff training cost like higher accuracy and longer training time is also the next step research.

References

1. Thomas, G. D., Ensemble methods in machine learning. In *Proc. of the First International Workshop on Multiple Classifier System (MCS 2000)*, 1–15, 2000.
2. Tsymbal, A., Pechenizkiy, M., and Cunningham, P., Diversity in search strategies for ensemble feature selection. *Inform. Fusion* 6:83–98, 2005.
3. Xing, E. P., et al., Feature selection for high-dimensional genomic microarray data. In *ICML'01: Proceedings of the Eighteenth International Conference on Machine Learning, 601–608*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2001.
4. Mitchell, T., *Machine learning*. McGraw-Hill, New York, 1997.
5. Yao, X., and Liu, Y., A new evolutionary system for evolving artificial neural networks. *IEEE Trans. Neural Netw.* 8:694–713, 1997.
6. Wang, Z., "Neuro-Fuzzy ensemble approach for microarray cancer gene expression data analysis," 2006 *International Symposium on Evolving Fuzzy Systems*, September, 2006.
7. Ding, C. H. Q., and Peng, H., Minimum redundancy feature selection from microarray gene expression data. In CSB, 523–529, 2003.
8. UC Irvine Machine Learning Repository http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Original%29
9. Mangasarian, O. L., and Wolberg, W. H., Cancer diagnosis via linear programming. *SIAM News* 23(5):1–18, 1990.
10. Quinlan, J. R., Improved use of continuous attributes in C4.5. *J. Artif. Intell. Res.* 4:7–90, 1996.
11. Sharkey, A. J. C., Sharkey, N. E., et al., Adapting an ensemble approach for the diagnosis of breast cancer. Proceedings of ICANN, 1998.
12. Abonyi, J., and Szeifert, F., Supervised fuzzy clustering for the identification of fuzzy classifiers. *Pattern Recogn. Lett.* 24:2195–2207, 2003.
13. Martínez-Muñoz, G., and Suárez, A., Switching class labels to generate classification ensembles. *Pattern Recogn.* 38:1483–1494, 2005.
14. Şahan, S., Polat, K., Kodaz, H., et al., A new hybrid method based on fuzzy-artificial immune system and k-nn algorithm for breast cancer diagnosis. *Comput. Biol. Med.* 37(3):415–423, 2007.