Ensemble Methods for Improving Classifier Performance

Monalisa Panda, Debahuti Mishra and Sashikala Mishra

Abstract In this paper, ensemble methods for different base classifiers are proposed. An ensemble technique is a supervised learning algorithm that combines a group of classifiers in order to acquire an overall model with more exact decisions. The classifiers that are support vector machine (SVM), naive Bayes (NB), and back propagation neural network (BPNN) are trained and tested on different gene expression datasets using both random selection method and *k*-fold cross-validation method. Both binary-class and multi-class datasets are used for evaluation of effectiveness of the ensemble method. Various publicly available gene expression datasets have been used for experiments in order to find the accuracy and effectiveness of the ensemble technique. Performance of the different classification methods and ensemble methods has been compared by using the accuracy values. The results have shown that the accuracy for the gene expression datasets has been increased by using the ensemble methods.

Keywords Gene expression · Ensemble · SVM · NB · BPNN
K-fold cross-validation

1 Introduction

Micro-array data is now used in many fields of medical diagnosis that is used for the detection of breast cancer, lymphoma, leukemia, etc. In order to measure the changes in expression levels of huge number of genes, micro-array data is used.

M. Panda (✉)
Department of CSE, CAPGS, BPUT, Rourkela, India
e-mail: monalisapanda1989@gmail.com

D. Mishra · S. Mishra
Department of CSE, ITER, Siksha 'O' Anusandhan University, Bhubaneswar, India
e-mail: mishradebahuti@gmail.com

S. Mishra
e-mail: sashi.iter@gmail.com

M.S. Reddy et al. (eds.), *International Proceedings on Advances in Soft Computing, Intelligent Systems and Applications*, Advances in Intelligent Systems and Computing 628, https://doi.org/10.1007/978-981-10-5272-9_34

Classification is a supervised learning process used for predicting a class label to any unseen data on the basis of training set of data, whose class label is already known. Nowadays, many existing classifiers such as SVM, *k*-nearest neighbor, ANN, Bayesian classifier, decision tree, linear regression are present. Commonly, a single classification method is not sufficient enough to correctly identify the class level. An ensemble technique is a supervised learning algorithm technique which combines a group of models in order to obtain an overall model with more precise decisions [1]. The models prediction, classification performance is usually improved by using the ensemble techniques.

Hence instead of choosing just one model, if we combine the outputs of different models, then the risk of selection of a badly performing classifier can be reduced. Several ensemble methods are there like voting, bagging, boosting, Bayesian merging, stacking, distribution summation, Dempster–Shafer, density-based weighting [2, 3]. This work mainly contains various classification and ensemble strategies, the set of laws for selecting the reduced data from large data sets, the act of using different classification techniques, how the classification and ensemble technique can be applied over different gene expression data sets. Here, stacking is used as an ensemble technique; that is, it combines the decisions of the individual classifier by using majority voting fusion rule. Stacking is concerned with combining multiple classifiers obtained by using various learning algorithms on a particular data set [4].

Finally, a comparison is done among different base classifiers and ensemble methods, and it was found that the ensemble methods were demonstrated with much better performance.

The rest of the paper is organized as follows: the basic definition of classifier ensemble is described in Sect. 2. Section 3 depicts the model. Section 4 explains the general methods, concepts, and approaches that are used to find out the result. Section 5 describes the two different ensemble techniques that are used to improve the result. Through simulation on variety of datasets, the result of the proposed model is reported in Sect. 5.

2 Classifier Ensemble Analysis

Classification is prediction of a certain result based on a given input. A training set containing a set of attributes and the result, usually called goal or prediction is being processed in order to predict the result. Classification in other words is a data mining function that assigns items in a group to mark categories or classes. Generally Classification is a process of estimating to which of a set of examples a new example belongs to, on the basis of a training dataset, whose class label is already known [2]. The algorithm that implements this process is known as classifier, which is a mathematical function that maps a data to a category.

In general, the single classification technique is not sufficient enough to identify the class level properly. An ensemble is itself a supervised learning algorithm which combines a set of models in order to obtain a global model with more accurate and reliable decisions [1]. When more number of algorithms is used in a model it becomes expensive. Therefore, nowadays, the researchers are emphasizing on the ensemble techniques. These techniques use to reduce the error rate in classification tasks in comparison with single classifiers. Also, the amalgamation of various techniques to make a final conclusion makes the performance of the system more strong against the difficulties that each individual classifier may have on each data set. Ensemble is mainly done to improve the accuracy and efficiency of the classification system.

3 Proposed Model

As mentioned earlier, this work focuses on the second phase of the model, that is, classifier ensemble techniques. In phase one, random selection method is used for training and testing of data. Here, we have used three classifiers, namely naive Bayes, backpropagation neural network, and support vector machine. In the second phase, *k*-fold cross-validation technique is used to divide the data set into training and testing. The value of *k* depends on the data set. Then training and testing is done up to *k* times for all the classifiers iteratively and then classifier fusion technique that is Stacking and Majority Voting are used to combine the outputs of the individual classifiers (Fig. 1).

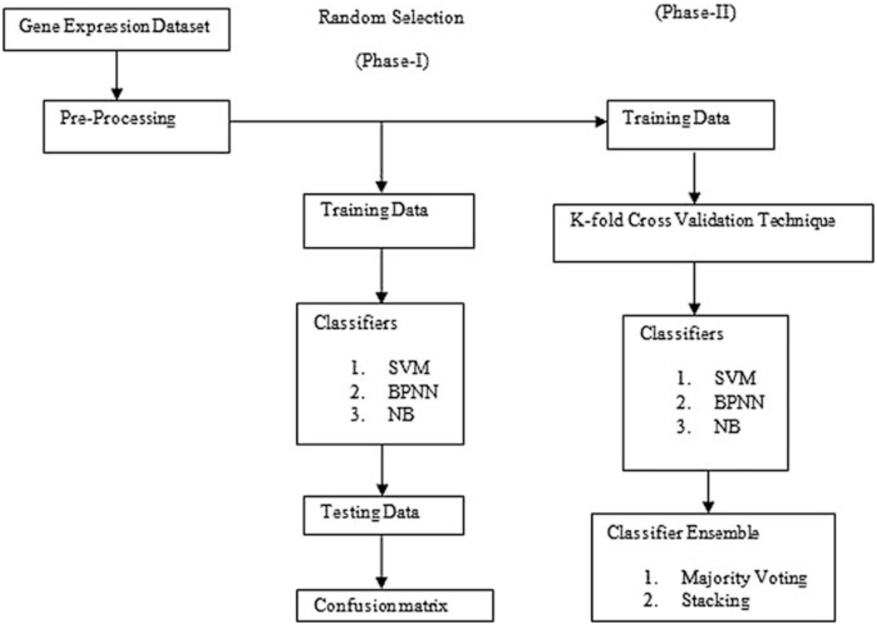

Fig. 1 Proposed model

4 Concepts, Methods, and Approaches

Initially, datasets need to be normalized. Data transformation such as normalization is a data preprocessing tool which is used in data mining system in order to remove the noisy data. An attribute of a dataset is getting normalized by scaling its values so that they fall within a specified range, such as 0.0–1.0 [5, 6]. Normalization is mostly useful for classification algorithms and clustering technique. Here, Min–Max normalization is used as a tool for preprocessing. Here, min A and max A are the minimum and maximum values of an attribute A. This technique can be calculated by using

$$V = (v - \min A)/(\max A - \min A). \tag{1}$$

After normalization, data reduction is done using PCA. The presence of large data sets can cause rigorous problems in an organization's decision support systems and database management systems. Micro-array data is high-dimensional data which can cause significant problems such as irrelevant genes, difficulty in constructing classifiers, and multiple missing gene expression values. In this paper, we have employed principal component analysis (PCA) as the feature reduction technique to extract the needful features, which can be used to train the classifiers. This feature reduced dataset is expected to provide a better classifier in terms of accuracy and efficiency. PCA is defined as a feature extraction method that transforms the data to a new coordinate system that is known as orthogonal linear transformation in such a way that by any projection of the data, the maximum variance comes to lie on the first coordinate that is known as the first principal component, then on the second coordinate lies the second largest variance and so on [7].

After feature reduction, the reduced data set is used for training by applying various classifiers like backpropagation neural network, support vector machine, and naive Bayes.

Backpropagation is learning or training algorithm rather than the network itself. A backpropagation learns by example. BPNN is a neural network learning algorithm that performs learning on multilayered feed-forward neural network. The training is completed by providing the input to the network, and the networks' weights are changed so that it will give us the required output for a particular input. In order to train the network we need to give the network examples of what we want the output (known as the Target) for a particular input. The weights are modified for each training data in order to reduce the error between the network's prediction and actual target value. Since the modifications are made in backward direction that is from the output layer to the hidden layer, hence, it is called backpropagation [8].

A naive Bayes classifier is defined as a probabilistic classifier that is based on applying Bayes theorem with some independence assumptions. In plain terms, a naive Bayes classifier assumes that the value of an individual feature is unrelated to the occurrence or lack of any other feature, provided with the class variable. An advantage of the naive Bayes classifier is that it only requires a small amount of training data to guess the parameters, that is, mean and variance of the variables that are necessary for the classification [9]. The Bayesian classification assumes a basic probabilistic model, and it allows to capture uncertainty about the model in a disciplined way by determining probabilities of the outcomes. It calculates explicit probabilities for hypothesis, and it is robust to noise in input data. Bayes theorem provides an approach to update the probability distribution of a variable based on information newly available by calculating the conditional distribution of the variable given the new information. The updated conditional probability distribution provides the new level of certainty about the variable. Posterior probability is calculated by updating the prior probability by using Bayes theorem. It uses the knowledge of prior events to predict the future events [10, 11]. Bayes theorem says:

$$P\left(\frac{\theta}{Y}\right) = P(\theta)^{*}P(Y/\theta)/P(Y), \tag{2}$$

where $P(\theta)$ and $P(Y)$ are the unconditional distributions of θ and Y. $P\left(\frac{\theta}{Y}\right)$ is the posterior distribution of θ.

$P(Y/\theta)$ is the likelihood function, and it measures how closely Y is distributed around θ.

SVM is used as a mapping function that transforms data in input space to data in feature space in a linearly separable manner [12, 13]. In machine learning, support vector machines (SVMs) are supervised learning models with associated learning algorithms, which analyze data and recognize patterns used for classification [14]. A support vector machine represents points in space, where the examples can be separated into distinct categories by a clear wide gap. Based on their category, new groups are being classified into one of those groups. In order to transform the original training data into a higher dimension, a nonlinear mapping is used. Support vector machines find a hyperplane which would be able to separate both the plane by retrieving the support vectors. SVM separates the hyperplane of class levels +1, −1 that is situated in maximum distance from both the positive and the negative samples. From both the negative and the positive pair, feature vectors are being extracted which are assigned with the class label of +1 and −1 to know whether the pair is a interacting or a non-interacting pair.

5 Classifier Ensemble Methods

An ensemble is itself a supervised learning algorithm which combines a set of models in order to obtain a global model with more accurate and reliable decisions [2, 15]. Classifier combination is one of the most frequently explored methods in data mining in the recent years. These techniques use to reduce the error rate in classification tasks in comparison with single classifiers. Therefore, nowadays, the researchers are emphasizing on the ensemble techniques. In this paper, majority voting and stacking are used on various gene expression datasets.

In majority voting, an unlabeled example is classified in accordance with the class that obtains the highest number of votes. It can be represented as follows:

$$\mathrm{Class}(X) = \arg\,\max_{c_i \in \mathrm{dom}(y)} \sum \forall k c_i = \arg\,\max_{c_i \in \mathrm{dom}(y) PM_k} \left(y = \frac{c_j}{x} \right), \quad (3)$$

where M_k denotes the classifier k and ${}_{PM_k}\left(y = \frac{c_j}{x}\right)$ denotes the probability of y obtaining the value of c at an instance x [16, 17].

Stacking is an ensemble method that is used for achieving the highest generalization accuracy. The reliability of the classifiers is judged on the basis of the meta-learner which learns from the outputs of the base learners. It uses the results of the base classifiers to produce a new record on which we need to apply a second learning algorithm [4]. This method allows us to maximize the utilization of the information contained in the training dataset. Normally to form a meta-learner training set, we divide the original training set into k disjoint subsets of equal size that is known as k-fold cross-validation technique [4, 18]. k will affect the overall accuracy boost and overall cost. The different base classifiers are trained and tested on different partitions of the training data. In the second level, again the classifiers are trained with the new class obtained from first level and the final accuracy is obtained. The results provided by this method were very good. The algorithm says as follows:

1. From the training set T, create k partitions from it and the cross-validation technique is used for all the base classifiers.
2. Machine learning is used to obtain second-level classifier.
3. A new class label is created and again uses the base classifiers to test the data and accuracy is found (Fig. 2).

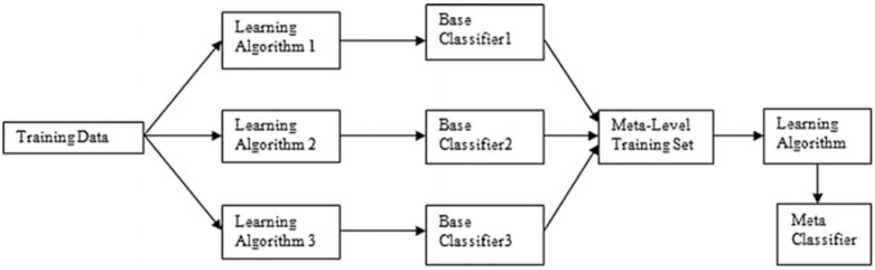

Fig. 2 Stacking technique

6 Results and Discussion

The set of experiments has been carried out using six datasets as shown in Table 1, such as breast cancer, lung cancer, iris, *E. coli*, yeast from UCI repository.

The proposed model has been tested with all the individual classifiers SVM, BPNN, NB, and the ensemble method that is stacking and majority voting for all five bench mark data sets as illustrated in Tables 2, 3 and 4. The threefold cross-validation test had been carried out, and the accuracy is measured. Entire algorithm is written and tested in MATLAB R2010a (Figs. 3, 4, 5, 6 and 7).

Table 1 Different datasets used for experimental evaluation

S. No.	Data set name	No. of instances	No. of attributes	No. of classes	References
1	Breast cancer	569	32	2	[19]
2	Lung cancer	32	56	4	[20]
3	Iris	150	4	3	[21]
4	*E. coli*	1484	9	10	[22]
5	Yeast	336	8	8	[23]

Table 2 Accuracy of different datasets using different classifiers using random selection method

S. No.	Data set name	Accuracy using BPNN	Accuracy using NB	Accuracy using SVM
1	Breast cancer	84.21	86.84	81.81
2	Lung cancer	84.28	85.71	86.84
3	Iris	83.33	85	84.84
4	*E. coli*	83.87	83.87	83.87
5	Yeast	86.80	87.81	86.80

Table 3 Accuracy of different datasets using stacking ensemble method

S. No.	Data set name	Accuracy using stacking BPNN	Accuracy using stacking NB	Accuracy using stacking SVM
1	Breast cancer	92.10	94.73	94.73
2	Lung cancer	97.14	100	94.28
3	Iris	93.33	95	95
4	*E. coli*	90.32	93.54	90.32
5	Yeast	95.17	93.44	95.86

Table 4 Accuracy of different datasets using majority voting ensemble method

S. No.	Data set name	Accuracy using majority voting	Execution time in s	Memory occupied in KB
1	Breast cancer	100	11.72	56.6
2	Lung cancer	98.57	12.70	28.6
3	Iris	100	11.86	11.7
4	*E. coli*	90.32	16.97	25.1
5	Yeast	98.62	93.43	92.6

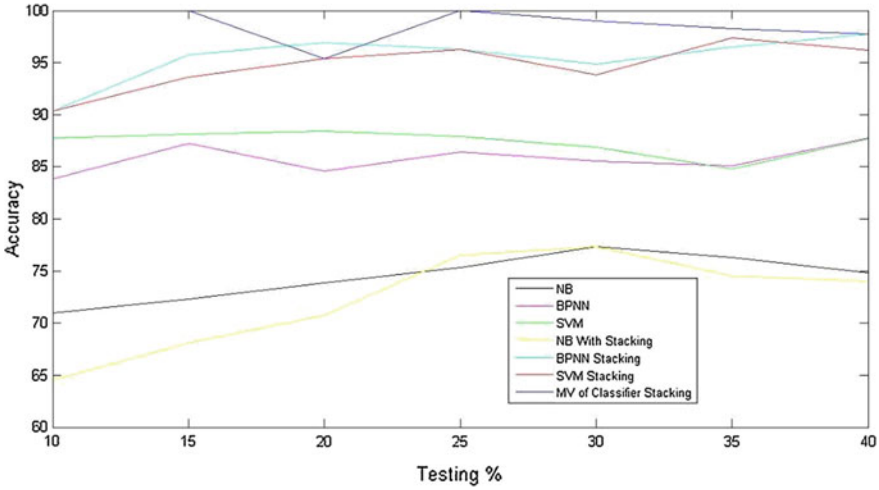

Fig. 3 Accuracy of classifiers and ensemble methods on breast cancer dataset

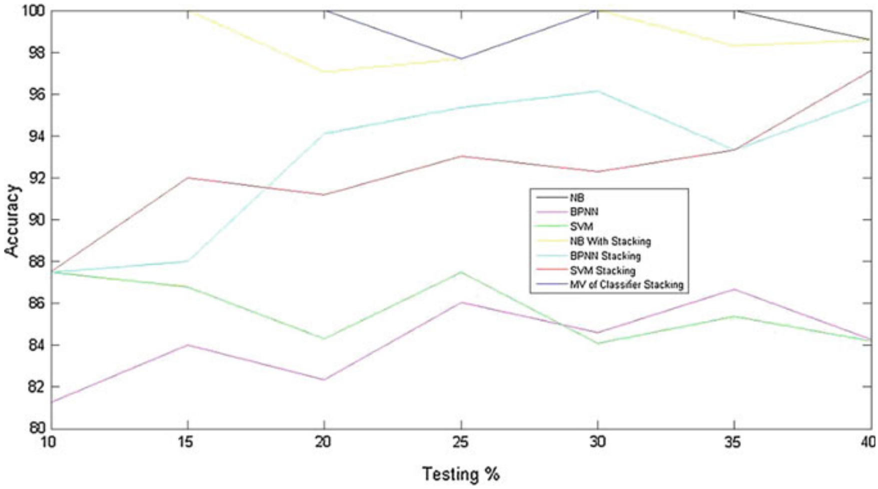

Fig. 4 Accuracy of classifiers and ensemble methods on lung cancer dataset

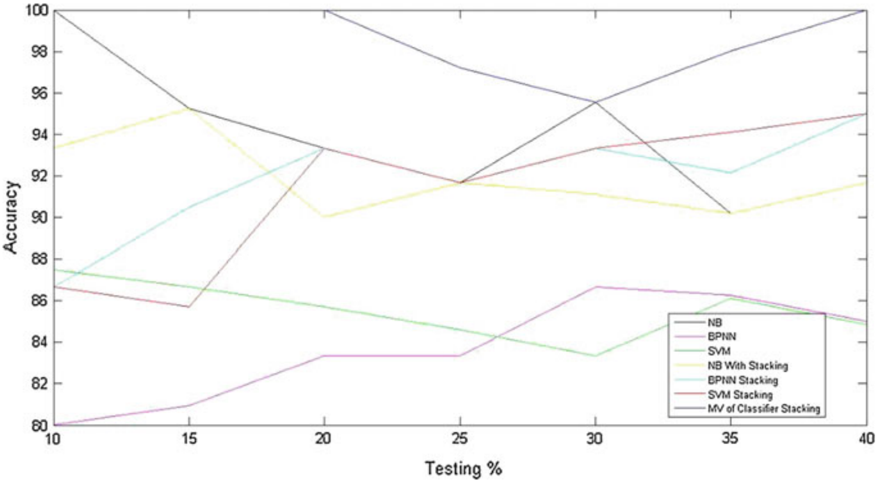

Fig. 5 Accuracy of classifiers and ensemble methods on iris dataset

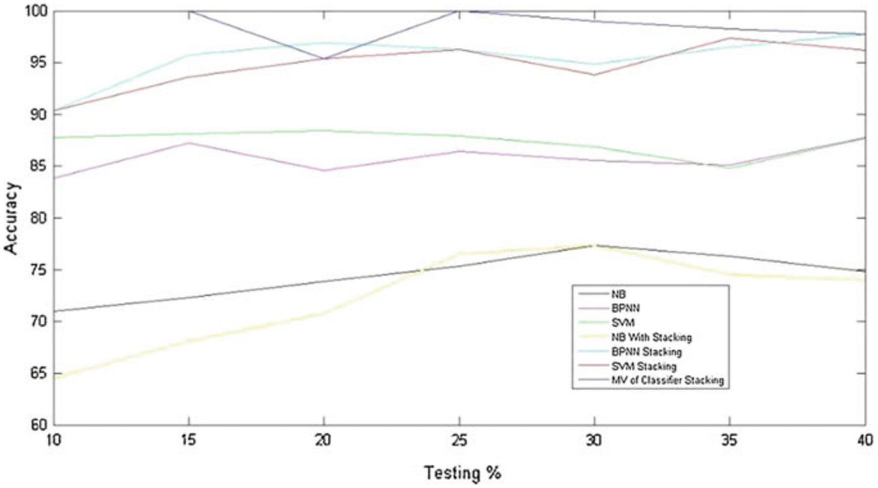

Fig. 6 Accuracy of classifiers and ensemble methods on *E. coli* dataset

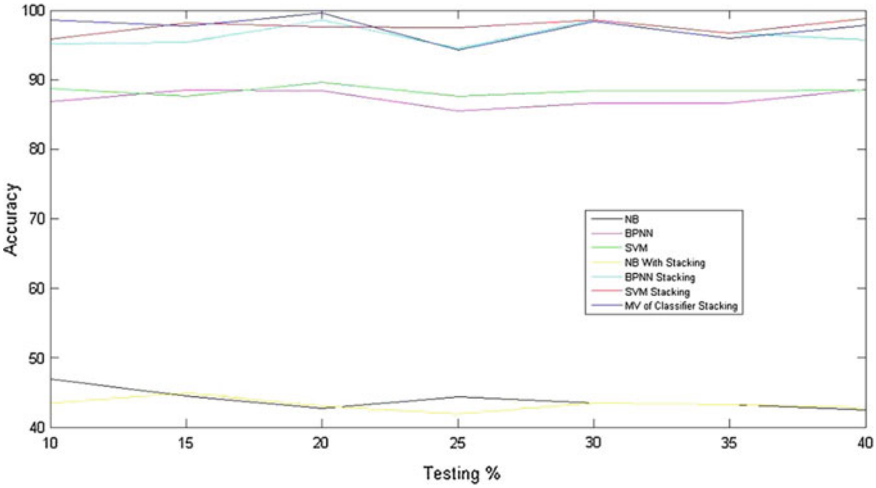

Fig. 7 Accuracy of classifiers and ensemble methods on yeast dataset

7 Conclusion

A comparative study is done between different classifiers and ensemble technique and are trained and tested on various publicly available gene expression datasets. Performance of the different classification methods and ensemble methods has been compared by using the accuracy values. The above ensemble methods that have been used for gene expression data set show that it achieves higher accuracy than all the other individual classifiers. The method also takes less computational time and space than the others.

Further, the accuracy of the ensemble technique can be enhanced much more by adding some optimization technique to the ensemble method.

References

1. Kun, M. 2013. A Vision-Based Hybrid Method for Eye Detection and Tracking. *International Journal of Security and Its Applications.*
2. Rokach, L. 2010. *Ensemble Methods in Supervised Learning*, vol. 33, 1–33. Springer.
3. Rokach, L. 2005. Ensemble Methods for Classifiers. *Data Mining and Knowledge Discovery Handbook*, Springer, US, 957–980.
4. Enriquez, F., F.L. Cruz, F. Javier Ortega, C.G. Vallego, and J.A. Troyano. 2013. A Comparative Study of Combination Applied to NLP Tasks. *Information Fusion* 14: 255–267.
5. Zhan, G.P. 2000. Neural Networks for Classification: A Survey. *IEEE Transactions on Systems, Man and Cybernetics-Part. C: Applications and Reviews* 30 (4): 451–446.

6. Ziadduin, S., and M.N. Dailey. 2008. Iris Recognition Performance Enhancement Using Weighted Majority Voting. *15th IEEE Inter-National Conference on Image Processing,* 227–280.
7. Isa, S.M., M. Ivan Fanany, W. Jatmiko, and A. Murni Arymurthy. 2011. Sleep Apnea Detection from ECG Signal: Analysis on Optimal Features. In *Principal Components and Nonlinearity, 5th International Conference on Bioinformatics and Biomedical Engineering.*
8. Kittler, J., M. Hatef, R.P.W. Duin, and J. Matas. 1998. On Combining Classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 2 (3): 226–239.
9. Kim, Seoyoung, and Y. Kim. 2012. Application-Specific Cloud Provisioning Model Using Job Profiles Analysis. In *IEEE 14th Conference on High Performance Computing and Communication and IEEE 9th International Conference on Embedded Software and Systems.*
10. Luo, L., E.F. Wood, and M. Pan. 2007. Bayesian Merging of Multiple Climate Model Forecasts for Seasonal Hydrological Predictions. *Journal of Geophysical Research* 112: 1–13.
11. Ajitha, P., and G. Gunasekaran. 2014. Semantic Based Intuitive Topic Search Engine. *International Review on Computers and Software*.
12. Chen, Z., J. Li, L. Wei, W. Xu, and Y. Shi. 2011. Multiple-Kernel SVM Based Multiple-Task Oriented Data Mining System for Gene Expression Data Analysis. *Expert Systems with Applications* 38: 12151–12159.
13. Hansen, L., and P. Salamon. 1990. Neural Network Ensembles. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 12: 993–1001.
14. Rokach. 2014. Decision Forests, Series in Machine Perception and Artificial Intelligence.
15. Helman, Paul, Robert Vero Susan, R. Atlas, and Cheryl Will-man. 2004. A Bayesian Network Classification Methodology for Gene Expression Data. *Journal of Computational Biology* 11 (4): 581–615.
16. Tsiliki, G., and S. Kossida. 2011. Fusion Methodologies for Biomedical Data. *Journal of Proteomics* 74: 2774–2785.
17. Kapp, M.N., R. Sabourin, and P. Maupin. 2012. A Dynamic Model Selection Strategy for Support Vector Machine Classifiers. *Applied Soft Computing* 12 (8): 2550–2565.
18. Dzeroski, S., and B. Zenko. 2004. Is Combining Classiers with Stacking Better Than Selecting the Best One? *Machine Learning* 54: 255–273.
19. Hong, Zi, and Jing-vu Yang. 1991. Optimal Discriminant Plane for a Small Number of Samples and Design Method of Classifier on the Plane. *Pattern Recognition* 24 (4): 317–324.
20. http://archieve.ics.uci.edu/ml/datasets/iris,2000-07-11.
21. http://archieve.ics.uci.edu/ml/datasets/yeast+dataset,1997-06-06.
22. http://archieve.ics.uci.edu/ml/datasets/ecoli,1997-06-06.
23. Seeja, K.R., and Shweta. 2011. Microarray Data Classification Using Support Vector Machine. *International Journal on Biometric and Bioinformatics* 5 (1): 10–15.
24. Shah, C., and A.G. Jivani. 2013. Comparison of Data Mining Classification Algorithms for Breast Cancer Prediction. In *Proceedings. 4th International Conference on Computing, Communication and Net-working Technologies*, 1–4.
25. ReboiroJato, M., F. Diaz, D. Glez-Pena, and F. Fdez-Riverola. 2014. A Novel Ensemble of Classifiers That Use Biological Relevant Gene Sets for Micro-array Classification. *Applied Soft Computing* 17: 117–126.
26. Opitz, D., and R. Maclin. 1999. Popular Ensemble Methods: An Empirical Study. *Journal of Artificial Intelligence Research* 11: 169–198.
27. Morrison, D., and L.C. De Silva. 2007. Voting Assembles of Spoken a ECT Classification. *Journal of Network and Computer Applications* 30: 1356–1365.
28. AliBagheri, M., Q. Gao, and S. Escalera. 2013. Logo Recognition Based on the Dempster-Shafer Fusion of Multiple Classifiers. *Advances in Artificial Intelligence Lecture Notes in Computer Science* 7884: 1–12.
29. Sohn, S.Y., and S. Ho Lee. 2003. Data Fusion, Ensemble and Clustering to Improve the Classification Accuracy for the Severity of Road Track Accidents in Korea. *Safety Science* 41: 1–14.

30. Hanczar, B., and A. BarHen. 2012. A New Measure of Classifier Performance for Gene Expression Data. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 9 (5): 1379–1386.
31. Tong, M., K. Hong Liu, C. Xu, and W. Ju. 2013. An Ensemble of SVM Classifiers Based on Gene Pairs. *Computers in Biology and Medicine* 43: 729–737.
32. Liu, H., L. Liu, and H. Zhang. 2010. Ensemble Gene Selection by Grouping for Microarray Data Classification. *Journal of Biomedical Informatics* 43: 81–87.
33. Reboiro-Jato, M., F. Diaz, D. Glez-Pena, and F. Fdez-Riverola. 2014. A Novel Ensemble of Classifiers That Use Biological Relevant Gene Sets for Microarray Classification. *Applied Soft Computing* 17: 117–126.
34. Nanni, L., and A. Lumini. 2007. Ensemblator: An Ensemble of Classifiers for Reliable Classification of Biological Data. *Pattern Recognition Letters* 28 (5): 622–630.
35. Lee, J., M. Park, and S. Song. 2005. An Extensive Comparison of Recent Classification Tools Applied to Microarray Data. *Computational Statistics and Data Analysis* 48 (4): 869–885.
36. Boulesteix, A., C. Strobl, T. Augustin, and M. Daumer. 2008. Evaluating Microarray Based Classifiers: An Overview. *Cancer Informatics* 6: 77–97.
37. Xu, L., A. Krzyzak and C.Y. Suen. 1992. Methods of Combining Multiple Classifiers and Their Applications to Handwriting Recognition. *IEEE Transactions on Systems, Man and Cybernetics* 22 (3): 418–435.
38. Chen, M.S., J. Han, and P.S. Yu. 1996. Data Mining: An Overview from a Database Perspective. *IEEE Transactions on Knowledge and Data Engineering* 8: 866–883.
39. Han, J., and M. Kamber. 2001. *Data Mining, Concepts and Techniques,* 67–120. Morgann Kaufmann Publishers.
40. Ester, M., H.P. Kriegel, J. Sander, and X. Xu. 1996. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In *Proceedings of 2nd International Conference on Knowledge Discovery and Data Mining*, vol. 96, 226–231.