A Novel Approach for Generating Clustered Based Ensemble of Classifiers

Mohammad Raihanul Islam, Md. Mustafizur Rahman, Asif Salekin, and Ahmed Shayer Andalib

***Abstract*—In this paper, we have presented a novel concept for constructing ensemble of classifiers. Here, we have considered a situation where data is available over the period of time. If enough remotely located data points are available at the classification system, the current system may not cope with newer data instances. In that case, existing settings or parameters of the classifiers need to be modified to act properly on newer instances. In this paper, we have presented a general technique for detecting enough remotely located data points arrived at the classifiers so that existing classification model can longer suitable for the new situation and proposed a change of settings to cope with the newer situation. We have performed detail analysis of our approach. Our approach has showed satisfactory results in dynamic environments.**

***Index Terms*—Neural networks, ensemble of classifiers, k-means clustering, chunk of data, identical classifier.**

I. Introduction

In the real world scenario, it is a common phenomenon that, all the data of a system are not available at the beginning. Data can be available throughout a definite period of time. It is very common in cases like weather or financial data. In this type of case, we have to design our classification model in such way that, it can adapt itself with the *newer* situation throughout the passage of time. The *newer* situations may include introduction of diverged data points or points located far from the current instances (remotely located) on the dataset or the introduction of newer classes. In both cases, the classifier has to adjust itself, so that it can learn the decision boundary of the newer classes without forgetting the previous knowledge.

An ensemble of classifiers can be defined as a set of classifiers which can be trained to learn the decision boundaries on training data. After the end of the training their decisions on the test examples are combined to obtain the final decision for the test data. The ensemble of classifications can also be defined as a group of classifiers or multiple classification system. There has been a significant amount of literature has been presented on generating ensemble of classifiers [1], [2]. An important characteristic of the ensemble of classifiers is to obtain the decisions of the classifiers in such way that, they differ from each other on labeling the class on identical pattern. This term is known as *diversity*. There has been some notable research works regarding diversity are presented in [3], [4].

There are various strategies to construct ensemble of classifiers. One of the popular approaches to generate ensemble of classifiers is by clustering. In this approach, the entire dataset can be divided into several disjoints subsets and for each subset a classifier can be trained. In this way, each of the classifiers become experts on the subsets and gives their decisions on the test patterns. For clustering the dataset, any clustering algorithm like *k-means* clustering can be used. Since we are considering that, data arrive at the classification system after a definite period of time, we can adjust the classification model to cope with the newer data by observing the position of data in Euclidean space. If the data is diverged from the current data points the parameters of the clustering algorithm can be adjusted to modify the classification model.

In this paper, we have presented an approach for constructing ensemble of classifiers using clustering where the data are available after a period of time. At first, we have detected that, whether there are sufficient numbers of data present in the newer instances so that, the existing classification system cannot handle the newer instances. So to cope with the newer instances the parameter of the ensemble system is changed. We have presented our approach as Ensemble of Classifiers using Clustering (ECC). We have experimented on our approach by varying the different distance metrics for clustering algorithm. We also have investigated the effect of varying the clustered parameters.

Our paper is organized as follows: In Section II we have presented some of the previous works regarding ensemble of classifiers. We describe our idea and approach in Section III. In Section IV we have described our experimental settings and have presented our results. We give our concluding remarks in Section V.

II. Background and Related Works

In the recent years, a significant amount of research has been directed to investigate the characteristics and the approach for generating ensemble of classifiers. Generally it can be observed that, these research works has been focused on two types of topics: the construction of the base classifiers and combining the decisions of the classifiers to reach the final verdict. In the construction method either identical classifiers or non-identical classifiers can be used.

There has also been some works regarding fusing the decisions of the classifiers [5], [6]. Here the final class label for the test example can be calculated using majority voting, weighted majority voting, Borda count etc. Also some other methods have been implemented for combing the decisions such as Dempster-Shafer combination, decision template [7].

Constructing the ensemble of classifiers using clustering

Manuscript received November 16, 2012; revised January 8, 2013.

The authors are with Bangladesh University of Engineering and Technology, Dhaka, Bangladesh (e-mail: kash_shaf91@yahoo.com).

can be found in [3], [8]. In [3], the authors introduce a layer based concept to generate ensemble of classifiers. In their approach, they have partitioned the data into using k-means clustering. A separate neural network is trained for each of the clustered datasets. They changed the parameters of the clustering algorithm, so that data points were grouped into different clusters each time. This way diverse decision on identical data is obtained so does the diversity.

Another method of constructing the ensemble of classifiers is to change the parameter of the classifiers to achieve diversity. One of the works regarding this idea is described in [9]. The experiment has been carried on NASA satellite images. Another method of constructing the ensemble of classifiers is to divide the feature space into different subsets and neural networks are trained on these subsets. The experiment is performed to recognize volcanoes on Venus [10]. In [1], authors have proposed an idea to construct ensemble of classifiers by manipulating the training examples. Here different training data are fed to train different training examples at different iteration. There is a weight for each training examples. If an instance is misclassified, then its weight is manipulated so that, it tends to be trained by the classifiers. Another popular technique for constructing ensemble of classifiers is commonly known as boosting. A generalized version of boosting known as AdaBoost is presented in [11].

Now a substantial amount of research directives have been concentrated on improving the performance of the ensemble of classifiers by applying different initiatives and reducing bias and variances [12]. Readers are suggested [3], [13] for review.

III. Novel Approach for Ensemble of Classifiers

In this Section, we describe our approach in detail. First we describe the basic principle of our approach. Next we discuss about the combining technique which is implemented in our approach.

A. Fundamental Concept

Our proposed approach is based on the concept of clustering. First we consider that, data are available to the classifiers continuously after a definite period of time. The entire data points can be partitioned by applying any clustering algorithm on them. In our approach, we have implemented k-means clustering algorithm. As a result, the data points are divided into disjoint subsets. Then any classifier like neural network can be trained on each of the clustered data points and thus each neural network become an expert on the clustered data.

Now consider that, a new chunk of data is available. Then we can measure the angle between each new data point with respect to the center of the cluster which is outside the range of any cluster exiting at the Euclidean space. The scenario has been depicted on Fig. 1. Here, a hypothetical example of scenario is presented. A sample cluster is shown with Grey background with center O. The newly arrived data points which are outside the range of cluster are also shown in small black circles. Now we measure the angle between the data points with respect to center of the cluster O. Now we can see from the figure that, data points close to each other have smaller angles between them, whereas remote data points have comparatively larger angles. So if we measure all the angles between each of the newer data points and see that, a definite number of them (α of the all the angles, α is number between 0 and 1) are greater than a certain threshold, then we can conclude that, a substantial number of diverged data have been introduced in the dataset. At this point, the existing classification model can no longer compatible with the new situation, so we change the parameter of the clustering algorithm to refine the classification model. As a result, changing the parameter settings will enable the classifiers to draw precise decision boundary between the diverged points.

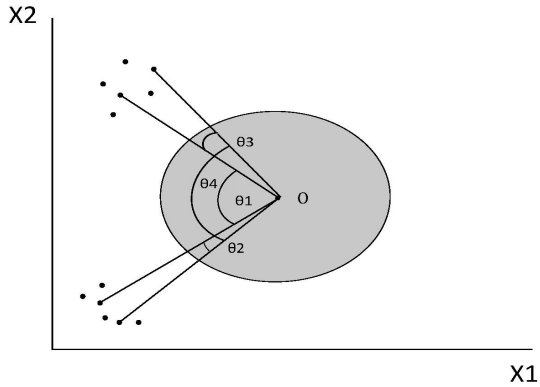

Fig. 1. A hypothetical example of the Euclidean space with arrival of new data points (Before applying K-means)

In our approach, we have increased the number of clusters in k-means clustering when the change in parameter settings is required. In Fig. 2, possible picture of Euclidean space after applying K-means clustering with increasing the value of number of cluster in K-means clustering is shown. Here, the previous cluster boundary is shown in dashed line while the newer boundary after applying K-means in continuous line.

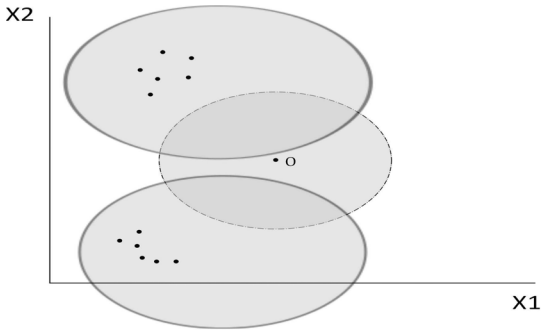

Fig. 2. A hypothetical example of the Euclidean space with arrival of new data points (After applying K-means)

The reason behind increasing the number cluster is that, if the number of clusters is remained same the existing cluster settings will not suitable for drawing a precise decision boundary or the classifiers will not be locally expert on the data points. This can be explained from Fig. 3, 4, 5 and 6.

In Fig. 3, we can see two clusters containing two classes (Black for one class and Grey for another). When new data are available, the scenario is shown in Fig. 4. If the number of cluster is not changed then, a probable case of cluster is shown in Fig. 5. We can get a clearer picture by looking at the figures and the necessity to increase the number of clusters. We will describe this step by step.

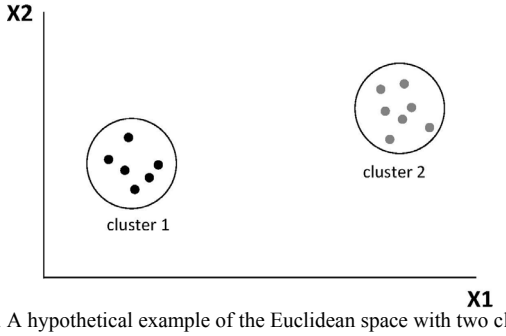

Fig. 3. A hypothetical example of the Euclidean space with two classes

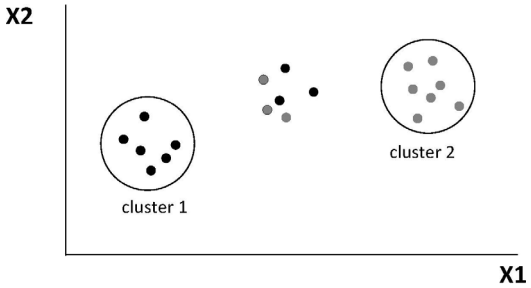

Fig. 4. A hypothetical example of the Euclidean space with arrival of new data points

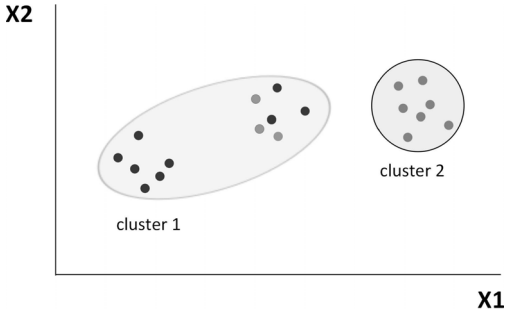

Fig. 5. A possible scenario of Euclidean space if number of cluster is not changed

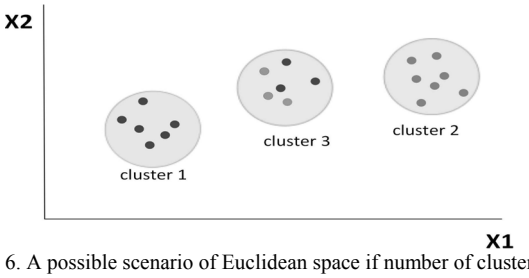

Fig. 6. A possible scenario of Euclidean space if number of cluster is increased

We can see that, since the new data are diverged, separating two classes in *cluster1* is difficult for the neural network trained on the data of that cluster. If we increase the number of clusters, then three clusters will form and three separate neural networks will be trained on each of the clustered data. The scenario is depicted in Fig. 6. Now the class boundary for each cluster will be more precise than it is before when the number of cluster the remained the same.

B. Fusing the Decision of the Classifier

One of the major elements of ensemble of classifiers is to implement a method to combine the decisions of the classifiers. There are many strategies presented in the literature. For our approach, we have applied *weighted majority voting*.

Consider a classification system, where there are T classifiers and C classes are available. Suppose the classifiers are *1,...,T* and the classes be *1,...,C*. Additionally assume that, the decision of the classifiers are define as $d_{t,j} \in \{0,1\}$ and if the t^{th} classifiers selects class ω_j, then $d_{t,j}=1$, otherwise $d_{t,j}=0$. Moreover, assume that, we have a mechanism to estimate the performance of the classifiers. Suppose a weight w_t is assigned to the classifier c_t with proportion to its evaluation. The weighted majority voting will choose class *J* if,

$$\sum_{t=1}^{T} w_t d_{t,J} = \max_{j=1}^{C} \sum_{t=1}^{T} w_t d_{t,J} \quad (1)$$

In our method, the weight of the classifiers is set with the inverse proportion of the Euclidean distance between the center of cluster of the data points with which the neural network has been trained and the test instance. The inverse proportion of the distance has been chosen since closer the data points to a certain cluster; the more probability is that, it will be rightly classified with the neural networks trained with the data points of that cluster. The overall procedure of our algorithm is shown in Fig. 7.

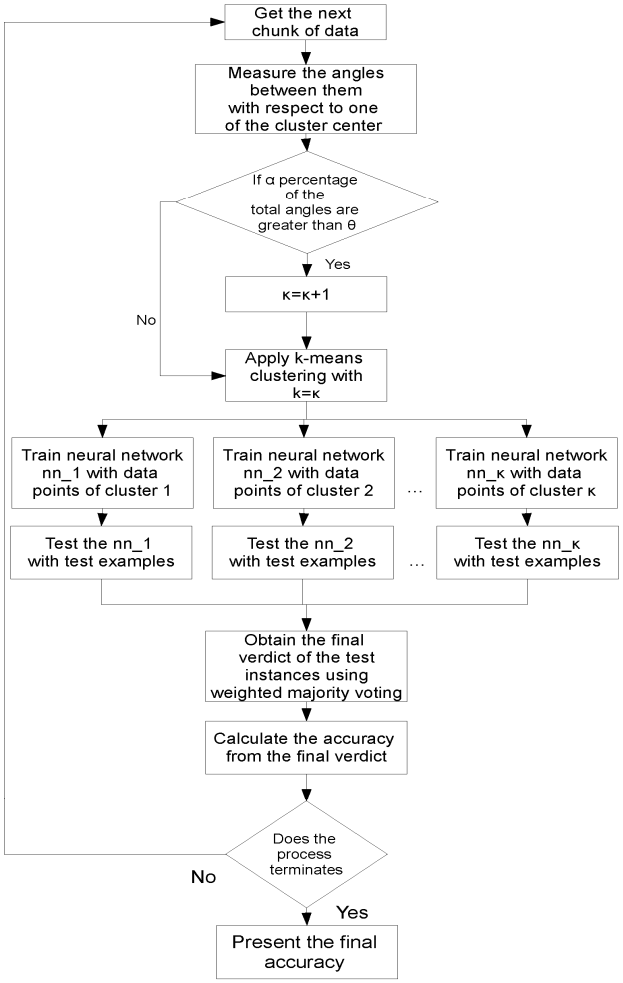

Fig. 7. Flow Chart for Total Procedure

IV. Experimental Setup and Results

In this Section, we present the detail description of our approach. First, we describe our experimental platforms. Next, we discuss the parameter settings and give a brief

description of the dataset we used. Later, we explain our instigation and discuss about the results.

We have conducted extensive experiments to estimate the performance of our algorithm. Our proposed algorithm has been tested with various datasets of UCI Machine Learning Repository. The dataset we used in our experiment is briefly described in Table I. The experiment has been carried on a PC with core *i*3 processor and 4GB RAM. The algorithm has been implemented on Matlab 2012.

TABLE I: DATASETS USED IN EXPERIMENTS

Dataset Name	Number of Instances	Number of Attributes	Number of Classes
Breast Cancer	699	9	2
Sonar	208	60	2
Transfusion	748	13	3
Wine	178	13	3
Satellite	6435	36	6
Pendigits	10992	16	10
Spam	4601	57	2

A. Parameters Settings

We have applied multi-layer neural network with back-propagation learning for training the clustered data. The number of hidden layer is 10 and the training iteration is 1000. The K-means algorithm has been implemented with difference distance metric which will be described shortly. K-means clustering starts with random initial points and the number of iteration for clustering was set to 5. In our experiment, the value of κ is set to 1.

Since the aim of the experiment is to classify data for dynamic environments, the parameters of the neural network are set as minimal as possible to perform classification task swiftly.

To explore the wide range of our approach, we have experimented on our algorithm with different parameters settings. The distance measurements we use in our experiment are described below:

Euclidean Distance: Suppose **x** and **y** are two n-dimensional vectors in Euclidean space, then the Euclidean distance between **x** and **y** is:

$$d(\mathbf{x},\mathbf{y}) = \sqrt{\sum_{k=1}^{n}(x_k - y_k)^2} \tag{2}$$

Absolute Distance: The absolute distance between **x** and **y** is:

$$a(\mathbf{x},\mathbf{y}) = \sum_{k=1}^{n} |x_k - y_k| \tag{3}$$

We also have varied the size of the data chunk. The range of chunk size is differed from 5% to 20% of total dataset. We have also changed the value of α. The value of was set 1%, 5%, 10% and 25%. Recall that, α was the percentage of total new chunk of data exceeding which cause the increasing of number of clusters. The threshold of the angle for increasing the number of cluster is set to 90° throughout the experiment. To measure the degree between the two data points with respect to a center of cluster the following equation is implemented:

$$\Theta = \cos^{-1}\frac{A.B}{|A||B|} \tag{4}$$

where *A* and *B* are the vectors drawn between the data points and the center of the cluster.

In our experiments, we have choose these distance metrics because these distance metric are the most popular. More distance metric can be implemented in practical purposes to fulfill specific criteria.

B. Results

We have shown the accuracy of the classifiers in Fig. 8. Here we have shown the accuracy with respect to the number of clusters. Since in our algorithm, the number of cluster is changed dynamically the accuracy shown at a certain cluster means that, when that number of clusters exists on the dataset. For example, the accuracy of cluster 1 in Breast Cancer Dataset is 90% means, when the number of cluster in that dataset is 1, the accuracy is 95%. The accuracy is 92% means that, when the number of cluster is 2, the accuracy is 92%.

From the Fig. 8, we can see that, in spite of the dynamic nature of the algorithm, the accuracy is high in all the cases. Only in the Transfusion Dataset some of the accuracy of the cluster is below 80%. Another important obeservation is that, when many instances are available to the algorithm tends to produce better results. Another notable points that can be observed that, the accuracy of the larger datasets increases as the number of cluster increases, which can be seen in Spam, Satellite and Pendigits; but in the smaller datasets the accuracy pattern is rather erratic.

Some regular pattern can be seen in Sonar dataset in which the number of instances is relatively small. The accuracy can be improved by applying neural networks with large number of hidden layer and increasing the training iteration, but it will also require significant amount of computational resources. Moreover, from the Fig. 8 we can also see that, how many cluster is formed at the termination of processes. For example, from Pendigits Dataset we can see that, number of cluster formed in our algorithm is 10; whereas for Sonar dataset 6 clusters are formed. We can also observe that, the number of clusters is around 6 in smaller datasets on the other hand in larger dataset more clusters tend to form.

TABLE II: COMPARISON TABLE

Dataset Name	Accuracy		
	ECC	AdaBoost	Bagging
Breast Cancer	85.3	70.28	64.69
Sonar	70.17	81.73	60.57
Transfusion	79.25	78.74	76.2
Wine	84.89	97.191	44.32
Satellite	86.64	87.6	11.67
Pendigits	93.84	93.76	97.8
Spam	91.62	91.43	64.11

We have compared our proposed techniques with existing models. We have compared our algorithms with AdaBoost and Bagging. The comparison is shown in Table II. From the table we can see that, the accuracy of ECC and AdaBoost are almost same. On the other hand, ECC shows much better accuracy over bagging in all respect. In this case, it should be considered that, ECC has been designed for handling dynamic environment while the others deal with static environment. So with promising accuracy our proposed ECC method can be highly suitable for dynamic environment.

V. Conclusion

In this paper, we have presented a novel concept for constructing ensemble of classifiers using clustering. We have presented detail descriptions and principles of our approach.

We have performed extensive experiments for our algorithms using real-life data from UCI Repository. From our experiments we have shown that, the result is more impressive in larger datasets. The overall accuracy is satisfactory considering the dynamic environment. We believe our method is highly suitable for scenarios where data are available after a definite period of time. We are very interested to investigate the possibilities of constructing this type of classification model suitable for real-life purpose.

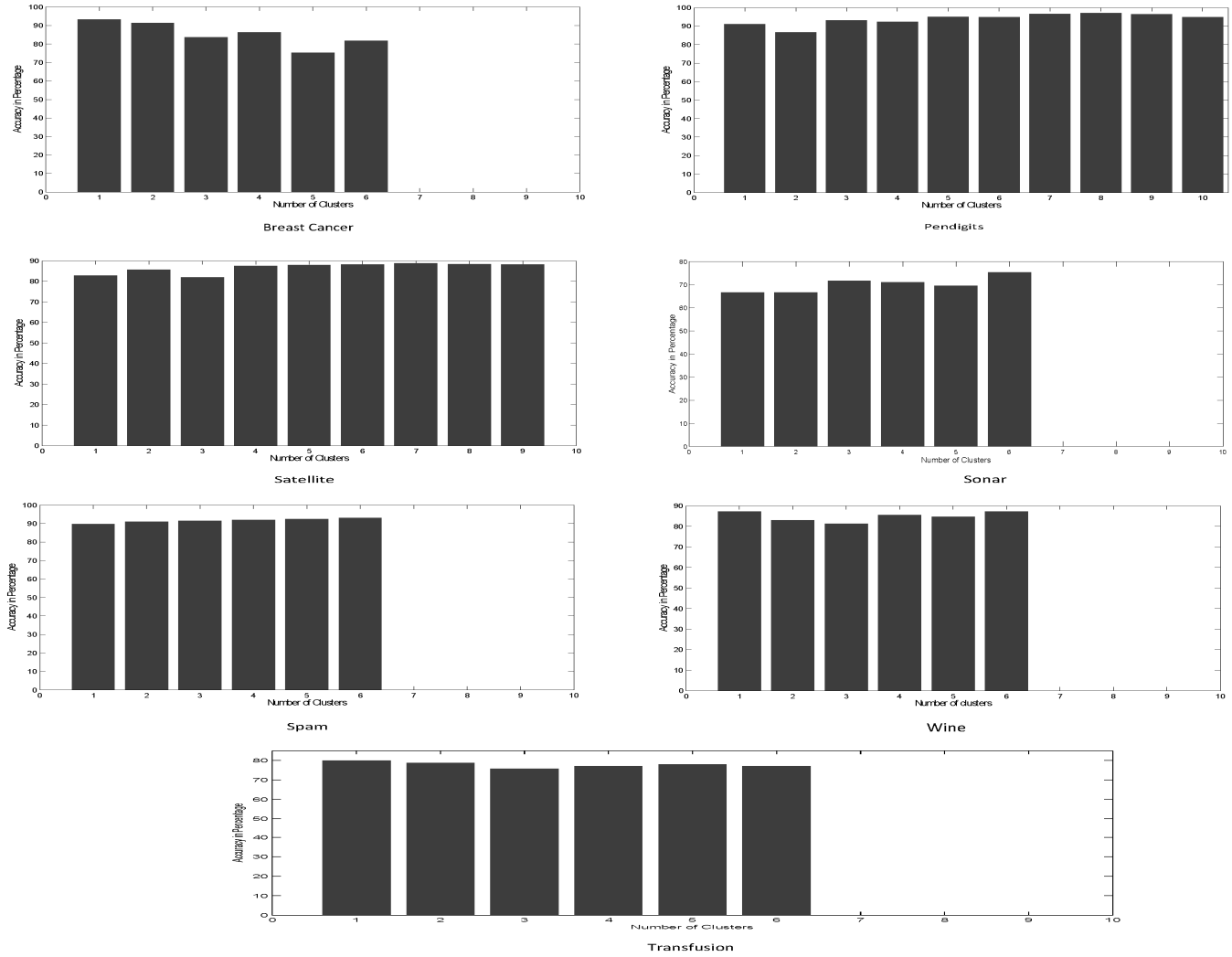

Fig. 8. Accuracy for different datasets

References

[1] R. Polikar, L. Udpa, S. S. Udpa, and V. Honavar, "Learn++: An Incremental Learning Algorithm for Supervised Neural Networks," *IEEE Trans. on System, Man, and Cybernatics*, vol. 31, no. 4, pp. 497-508, Nov. 2001.

[2] R. Polikar, "Ensemble based systems in decision making," *IEEE Circuits and Systems Magazine*, vol. 6, no. 3, pp. 21-45, 2001.

[3] A. Rahman and B. Verma, "Novel Layered Clustering-Based Approach for Generating Ensemble of Classifiers," *IEEE Transactions on Neural Networks,* vol.22, no.5, pp. 781-792, 2011.

[4] H. Zouari, L. Heutte, and Y. Lecourtier, "Controlling the diversity in classifier ensembles through a measure of agreement," *Pattern Recognition,* vol. 38, no. 11, pp. 2195-2199, Nov. 2005.

[5] J. Kittler, M. Hatef, R. P. W. Duin, and J. Matas, "On combining classifiers," *IEEE Transactions Pattern Analysis and Machine Intelligence*, vol. 20, no. 3, pp. 226–239, Mar. 1998.

[6] L. K. Hansen and P. Salmon, "Neural network ensembles," *IEEE Transactions Pattern Analysis and Machine Intelligence*, vol. 12, no. 10, pp. 993-1001, Oct. 1990.

[7] L. I. Kuncheva, J. C. Bezdek, and R. P. W. Duin, "Decision templates for multiple classifier fusion: An experimental comparison," *Pattern Recognition*, vol. 34, pp. 299-314, 2001.

[8] R. Neumayer, "Clustering based ensemble classification for spam filtering," in *Proceedings of the Seventh Workshop on Data Analysis*, 2006.

[9] T. Yamaguchi, K. J. Mackin, E. Nunohiro, J. G. Park, K. Hara, K. Matsushita, M. Ohshiro, and K. Yamasaki, "Artificial neural network ensemble-based land-cover classifiers using MODIS data," *Artificial life and Robotics*, vol. 13, no. 2, pp. 570-574, 2009.

[10] K. J. Cherkauer, "Human Expert-Level Performance on a Scientific Image Analysis Task by a System Using Combined Artificial Neural Networks," in National Conference on Artificial Intelligence, 1996, pp. 15-21.

[11] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *Journal of Computer and System Sciences*, vol. 55, no. 1, pp. 119-139, 1997.

[12] L. I. Kuncheva, *Combining Pattern Classifiers: Methods and Algorithms*. Wiley-Interscience, 2004.

[13] R. Polikar, "Bootstrap—Inspired techniques in computation intelligence," *IEEE Signal Processing Magazine*, vol. 24, no. 4, pp. 59-72, Jul. 2007.