



## International Journal of Emerging Technology and Advanced Engineering

Website: [www.ijetae.com](http://www.ijetae.com) (ISSN 2250-2459 (Online), An ISO 9001:2008 Certified Journal, Volume 3, Special Issue 2, January 2013)

National conference on Machine Intelligence Research and Advancement (NCMIRA, 12), INDIA.

# Improving the Classification Accuracy with Ensemble of Classifiers

A K Saxena<sup>1</sup>, Apoorva Mondal<sup>2</sup>, Itfaq Ahmad Mir<sup>3</sup>

<sup>1</sup>Professor, Dept of CSIT, <sup>2</sup>Student, B.Tech V Sem.; Guru Ghasidas Vishwavidyalaya, Bilaspur(CG)

<sup>3</sup>Itfaq Ahmad Mir, Prog Asstt (Computer), KVK (Pulwama), SKUAST-K, Srinagar, J&K

### Abstract

In this paper we investigate a recent technique for classification of datasets. One of the major factors to evaluate a classifier depends on how accurately it can classify unknown patterns. There are a number of classification algorithms, both supervised and unsupervised. In most cases, a single classifier is trained on a part of the dataset and tested on the remaining part of the same dataset. It is observed that a single classifier performing excellently for the particular part of a dataset produces poor classification accuracy when presented with another part of the same dataset. In this paper, an ensemble approach of classification is presented. Consequently, various parts of the dataset are trained using individual k-nearest neighbors (k-*nn*) classifiers. Using bagging and majority of voting techniques, the classification accuracy of test dataset is evaluated. Three benchmark datasets are used for empirical study of the scheme. After extensive experiments with three ensembles having different number of classifiers, three different values of *k* for three different sizes of training datasets, selected randomly on five trials, it is observed that the ensemble of classifiers produces better classification accuracies than does any individual classifier.

**Keywords:** Classification, Ensemble of classifiers, k-*nn* classifier

### I. INTRODUCTION

With the emergence of several data analysis techniques like data mining[1], the role of classification has become very important. In most of the cases, the large database is required to be grouped on the basis on certain criteria. Classification is commonly used to separate the database into different classes. Classification is one of the core challenging tasks [2] in data mining [3], pattern recognition [4], bioinformatics [5] etc [6]. The goal of classification [7] is to assign a new entity into a class from a pre-specified set of classes. The purpose of supervised learning is to classify patterns (also known as instances) into a set of categories which are also referred to as *classes* or *labels*. Commonly, the classification is based on classification models (classifiers) that are induced from an exemplary set of pre classified patterns. Alternatively, the classification utilizes knowledge that is supplied by an expert in the application domain. In a typical supervised learning setting, a set of instances, also referred to as a *training set* is given. The labels of the instances in the training set are known and the goal is to construct a model in order to label new instances. An algorithm which constructs the model is called *inducer* and an instance of an inducer for a specific training set is called a *classifier*.

The main idea behind the ensemble methodology is to weigh several individual classifiers, and combine them in order to obtain a classifier that outperforms every one of them[8]. In fact, human being tends to seek several opinions before making any important decision. We weigh the individual opinions, and combine them to reach our final decision [9]. Recently, Mikel Galar et al.[10] reported that class distribution, i.e., the proportion of instances belonging to each class in a data-set, plays a key role in classification. Imbalanced data-sets problem occurs when one class, usually the one that refers to the concept of interest (positive or minority class), is underrepresented in the data-set; in other words, the number of negative (majority) instances outnumbers the amount of positive class instances[11]. The primary benefit of using ensemble systems is the reduction of variance and increase in confidence of the decision. Due to many random variations in a given classifier model (different training data, different initialization, etc.), the decision obtained by any given classifier may vary substantially from one training trial to another—even if the model structure is kept constant. Then, combining the outputs of several such classifiers by, for example, averaging the output decisions, can reduce the risk of an unfortunate selection of a poorly performing classifier.



## International Journal of Emerging Technology and Advanced Engineering

Website: [www.ijetae.com](http://www.ijetae.com) (ISSN 2250-2459 (Online), An ISO 9001:2008 Certified Journal, Volume 3, Special Issue 2, January 2013)

### National conference on Machine Intelligence Research and Advancement (NCMIRA, 12), INDIA.

Another use of ensemble systems includes splitting large datasets into smaller and logical partitions, each used to train a separate classifier. This can be more efficient than using a single model to describe the entire data. The opposite problem, having too little data, can also be handled using ensemble systems, and this is where bootstrap-based ideas start surfacing: generate multiple classifiers, each trained on a different subset of the data, obtained through bootstrap resampling. While the history of ensemble systems can be traced back to some earlier studies such as [12,13], it is Schapire's 1990 paper[14] that is widely recognized as the seminal work on ensemble systems.

In this paper, an ensemble approach of classification is presented. Consequently, various parts of the dataset are trained using individual k-nearest neighbors (k-nn) classifiers. Using bagging and majority of voting techniques, the classification accuracy of test dataset is evaluated. Three benchmark datasets are used for empirical study of the scheme. After extensive experiments with three ensembles having different number of classifiers, three different values of k for three different sizes of training datasets, selected randomly on five trials, it is observed that the ensemble of classifiers produces better classification accuracies than does any individual classifier.

#### II. PROPOSED ENSEMBLE ALGORITHM

The ensemble of classifiers can be constructed with techniques outlined in [8,11]. Here, the proposed algorithm is presented keeping in mind the nature of supervised classifiers and three benchmark datasets used in the paper.

##### Algorithm

1. Divide the entire dataset D into two parts, training dataset,  $S_{tr}$  and testing dataset,  $S_{te}$ . Each part has got same number of features. Number of patterns in these two can be different but each pattern has got one class.
2. Decide the number of classifiers, S, for a training dataset to start with. Make equal partitions of  $S_{tr}$  such that all except the last part, will have  $S_{tr}/S$  patterns. The last part can have the remaining patterns if  $S_{tr}$  is not divisible by S. The ensemble will thus have S classifiers.
3. Call a classifier, k-nearest neighbor, k-nn used in this paper. Set a value for k.

4. Determine the classification accuracy, CA of each part of the training data against the same test data set  $S_{te}$ .
5. Run the experiment for at least 5 times by shuffling dataset D and changing  $S_{tr}$  and  $S_{te}$  every time to ensure proper diversity and representation of all classes in training and testing datasets. Find the average CA.
6. Take every pattern of  $S_{te}$  and pass it through all classifiers using bagging [8,11] and majority of voting techniques to find out it's class. This is also repeated five times, Calculate average CA of the ensemble ECA.
7. Change the value of k.
8. Change the value of S.
9. Change the size of training and test data.

The proposed algorithm can be viewed by Fig. 1. In this figure, as a typical example, five classifiers are used to be put in an ensemble. The parts of training data  $S_1...S_5$  are used for creating five classifiers  $C_1...C_5$ , one classifier for one part respectively. The CA of ensemble is shown by  $C_e$ .

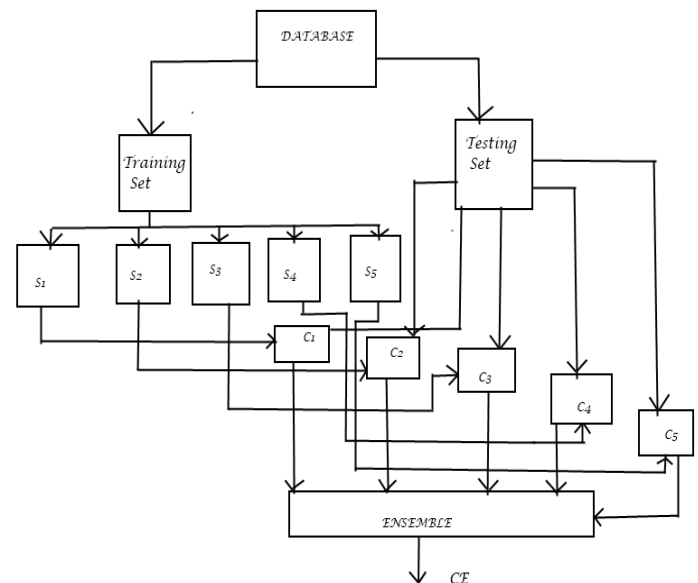


Fig.1.Representation of ensemble algorithm for number of classifiers, S=5



## International Journal of Emerging Technology and Advanced Engineering

Website: [www.ijetae.com](http://www.ijetae.com) (ISSN 2250-2459 (Online), An ISO 9001:2008 Certified Journal, Volume 3, Special Issue 2, January 2013)

### National conference on Machine Intelligence Research and Advancement (NCMIRA, 12), INDIA.

#### III. SUMMARY OF DATABASES

Table 1 summarizes the data sets used for the experiments. These are the bench mark data bases and can be found in [15].

**TABLE 1**  
Description of the Data Set Used.

Data Set	Total Patterns	Attributes	Classes	Patterns in Class1	Patterns in Class2	Patterns in Class3
Iris	150	4	3	50	50	50
Wine	178	13	3	59	71	48
Bupa	345	6	2	145	200	-

#### IV. EXPERIMENTS AND RESULTS

Proposed scheme was run on an i5 machine using MATLAB. The results obtained for the three databases are shown in Table 2 (A), 2(B) and 2(C) for Iris, Liver and Wine datasets respectively. In each of these tables, first column: training data size indicates the part (in percents) of the database which will be used for training only whereas the remaining part (100 – training data) will be used for testing. Three sizes for training have been used in the paper viz. 80%, 60% and 40%, to reflect attitude of the proposed scheme towards different parts of the data. The next column represents values of ‘k’, i.e. the k-th nearest neighbor from the testing data pattern. The measure of the distance is taken as Euclidean distance. Three values of ‘k’ ( 1, 3 and 5), have been used for all three datasets. To apply bagging, each training dataset is further divided into  $S$  number of sub sets. In the paper,  $S$  is set for three values: 5, 7 and 9. In other words, number of classifiers in ensemble will be 5, 7 and 9 for each of the three datasets. Thus for each dataset, a training part of the dataset (80/60/40 %), has  $S$  different subsets. For a typical training dataset with five folds or subsets

$$S_1 \cup S_2 \cup S_3 \cup S_4 \cup S_5 = S_{tr}$$

and

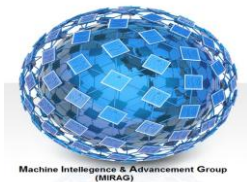
$$S_{tr} \cup S_{te} = D$$

Where  $S_{tr}$  and  $S_{te}$  stand for training and testing Dataset respectively, and  $D$  is the entire dataset. As a typical case, first experiment is conducted with  $S=5, k=1$  and training data size =80% of the total dataset. The testing data (20%) will remain as the unseen part of dataset. In this case, five classifiers operate on five subsets ( $S_1, \dots, S_5$ ) to produce five classification accuracies in percents with  $k=1$  with the same test dataset. The ensemble accuracy in percent, is determined as majority of opinions about a class by these five classifiers. Time in execution of the whole process is also recorded. This exercise is repeated for five times by shuffling randomly the dataset five times for diversity. Take the average classification accuracy  $ACA$ , average ensemble accuracy  $ECA$  and average time spent on five executions. This is shown by the first row of the first main sub column of Table with  $S=5$ . Similar exercise is repeated for  $S=7$  and 9. This completes row 1 of the table. The values of  $k$  are varied to 3 and 5. Then training data size is changed to 60% and 40% and exactly same procedure is adopted. All the entries are filled in Table 2(A).

On observing Tables 2(A),(B) and (C), it is noted that for iris data set, for  $S=5, k=1$ ,  $ACA =92.2$  is highest when individual classifiers are considered. In this case  $ECA$  is 93.3. For  $S=7, k=1$ ,  $ACA =93.2$  is highest for individuals, whereas  $ECA =100$ . For  $S=9, ACA=92.1, ECA =97.8$ . Thus it is noted that  $ECA$  is in each case is higher than  $ACA$ . For 60% training data,  $ACA =93.4$  for  $S=5, k=3$ , with  $ECA =95.0$ ; for  $S=7, ACA=90.3, k=1, ECA=94.4$  for  $k=3$ ; for  $S=9, k=1/3, ACA =87.8$  and  $ECA=96.3$ . It indicates that  $ECA$  is higher in each case under 6% training as well. For 40% training data,  $ACA=91.0, k=1, ECA=91.5$ ; for  $S=7, k=1, ACA=89.6, ECA=94.4, k=3$ ; for  $S=9, k=5, ACA=82.6, ECA=94.1$ . It is again noted that  $ECA$  is better in each case.

During certain experiments, it was seen that individual classifier for a certain dataset performed better but when presented with other dataset, the performance turned poor, whereas, ensembles produced better classifier accuracy than individual classifiers. Similar observations are seen in Tables 2(B) for Liver Data and 2(C) for Wine Data.

It is therefore noted in general that  $ECA$  produces higher CA than any individual classifier in each of the three datasets used for empirical studies.



## International Journal of Emerging Technology and Advanced Engineering

Website: www.ijetae.com (ISSN 2250-2459 (Online), An ISO 9001:2008 Certified Journal, Volume 3, Special Issue 2, January 2013)

National conference on Machine Intelligence Research and Advancement (NCMIRA, 12), INDIA.

TABLE 2

### COMPARISON OF PERFORMANCES OF ENSEMBLES WITH INDIVIDUAL CLASSIFIERS

Abbreviations used in Tables

TDS: Training Data Size in %; S: No. of Ensembles (Subsets of training data), CA: Classification Accuracies; k: No. of nearest neighbors =k; ACA: Average Classification Accuracy; ECA: Ensemble CA; T: Total time in training, testing with individual classifiers and that with ensembles

TABLE 2(A) – IRIS DATA

TDS	k	S=5			S=7			S=9		
		AC A	EC A	T	AC A	EC A	T	AC A	EC A	T
80	1	92.2	93.3	0.31	93.2	96.7	0.32	92.1	95.6	.31
	3	91.3	93.3	0.33	92.5	100	0.32	92.1	97.8	0.31
	5	82	93.3	0.33	92.9	96.7	0.33	82.2	94.4	0.31
60	1	90.7	93.7	0.31	90.3	92.2	0.33	87.8	96.3	0.33
	3	93.4	95.0	0.32	87.5	94.4	0.32	87.8	96.3	0.33
	5	88.6	91.1	0.32	83.6	92.2	0.36	66.5	89.2	0.36
40	1	91.0	91.5	0.32	89.6	94.1	.31	82.5	94.1	0.33
	3	78.7	89.6	0.35	73.7	94.4	0.35	82.6	94.1	0.33
	5	75.1	87.0	0.34	55.0	54.1	0.37	83.5	93.8	0.36

TABLE 2(B) - BUPA LIVER DATA

TDS	K	S=5			S=7			S=9		
		ACA	EC A	T	AC A	EC A	T	AC A	EC A	T
80	1	59.0	63.0	0.50	57.6	65.2	0.38	57.7	65.2	0.39
	3	60.6	64.7	0.41	59.6	63.8	0.43	57.0	63.4	0.40
	5	59.6	62.7	0.38	56.9	61.4	0.39	58.1	60.9	0.40
60	1	58.2	66.1	0.42	58.4	65.2	0.41	58.2	66.3	0.41
	3	58.3	63.9	0.42	57.7	62.5	0.42	57.4	63.0	0.42
	5	60.8	66.4	0.40	61.2	69.1	0.41	59.4	67.0	0.44
40	1	56.9	60.1	0.42	56.6	62.4	0.43	55.9	61.8	0.43
	3	60.3	67.6	0.42	58.5	61.7	0.42	57.2	64.2	0.44
	5	58.8	62.6	0.44	55.9	59.2	0.41	56.1	62.6	0.45

TABLE 2(C) – WINE DATA

TDS	K	S=5			S=7			S=9		
		AC A	EC A	T	AC A	EC A	T	AC A	EC A	T
80	1	72.4	75.9	0.34	67.1	74.3	0.35	65.5	69.4	0.35
	3	73.3	77.8	0.37	68.5	71.5	0.37	62.2	62.0	0.36
	5	68.3	69.4	0.37	68.3	73.0	0.37	68.5	72.2	0.39
60	1	67.2	68.3	0.34	65.2	69.0	0.34	65.9	70.4	0.35
	3	67.1	66.2	0.36	65.0	66.7	0.35	65.3	68.0	0.39
	5	69.0	70.0	0.39	61.0	67.1	0.39	65.1	67.6	0.40
40	1	67.1	69.5	0.34	65.1	72.7	0.34	60.8	64.0	0.36
	3	66.1	68.0	0.38	61.6	67.8	0.39	59.7	72.4	0.39
	5	67.0	66.7	0.40	55.6	59.3	0.38	53.8	67.3	0.40

### V. CONCLUSION

Classification is an essential component of various data analysis methods. It has extensive applications in several areas of science, engineering, technology, medical and social studies. In this paper, a recent yet important scheme of classification has been presented. For classification purpose, there have been different established classifiers which are reported in the literature from time to time with extensions. In this paper, it is attempted to claim that an individual classifier may not be able to predict the class of an unknown pattern correctly. On the other hand, if the multiple classifiers are trained for small databases, and they are combined in the form of an ensemble, then the prediction of the class by bagging of these classifiers in ensemble can be more accurate. To justify it, three benchmark datasets, iris, BUPA liver and wine have been used for empirical study. The training part of each dataset is varied into three cases viz. 80%, 60% and 40%. The number of classifiers are also varied with 5, 7 and 9 sub sets (or folds) of the training dataset. As a case study, k-nn classifiers are used in each case. The values of k are set to 1, 3 and 5 in all cases. Experiments were conducted to evaluate the classification accuracies under all such cases. In order to provide diversity in creating training and testing datasets, the entire experiment was iterated for 5 times for each case. Under such extensive environment of experiments, the results produced are stored in tables. From observation of these tables, it is noticed that the CA of each individual classifier is dominated by ensemble's CA. Thus it is concluded from these investigations that an ensemble is a good approach to determine the class of an unseen data pattern. The study can be extended to many other datasets. Moreover, other classifiers like neural network, fuzzy etc. can be included in the ensemble.

### REFERENCES

- [1] Kamber, M., Han, J., Data mining: Concepts and techniques, 2nd ed. CA: Morgan Kaufmann Publisher. San Francisco, 2006.
- [2] Misra, B.B., Dehuri, S., Dash, P.K., Panda, G., "Reduced Polynomial Neural Swarm Net for Classification Task in Data Mining", IEEE Congress on Evolutionary Computation, 2008b
- [3] Kosala, R., Blockeel, H., "Mining Research: A Survey". ACM SIGKDD Explorations. 2 (1), 2000, pp 1–15.
- [4] Duda, R.O., Hart, P.E., Stork, D.G., Pattern Classification. John Wiley and Sons (Asia) Pte. Ltd., 2001.
- [5] Baldi, P., Brunak, S., Bioinformatics: The Machine Learning Approach. MIT Press, Cambridge, MA., 1998.



## International Journal of Emerging Technology and Advanced Engineering

Website: [www.ijetae.com](http://www.ijetae.com) (ISSN 2250-2459 (Online), An ISO 9001:2008 Certified Journal, Volume 3, Special Issue 2, January 2013)

### National conference on Machine Intelligence Research and Advancement (NCMIRA, 12), INDIA.

- [6] Saxena, Patre, Dubey, "An Evolutionary Feature Selection Technique Using Polynomial Neural Network", IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 4, No 1, July 2011 ISSN (Online): pp. 1694-0814 [www.IJCSI.org](http://www.IJCSI.org).
- [7] Mitchel, T.M., Machine Learning. McGraw Hill, 1997.
- [8] Lior Rokach, "Ensemble-based classifiers", Artif Intell Rev pp. 33:1-39, Springer, 2010.
- [9] Polikar R, "Ensemble based systems in decision making", IEEE Circuits Syst Mag 6(3):pp. 21-45, 2006.
- [10] Mikel Galar et al. , "A Review on Ensembles for the Class Imbalance Problem: Bagging-, Boosting-, and Hybrid-Based Approaches", IEEE Transaction on Systems, Man and Cybernetics, Part-C, Applications and reviews, pp. 1-22, 2011.
- [11] Polikar, "Bootstrap methods", in IEEE Signal Processing Magazine, July 2007.
- [12] B.V. Dasarathy and B.V. Sheela, "Composite classifier system design: Concepts and methodology," Proc. IEEE, vol. 67, no. 5, pp. 708-713, 1979.
- [13] L.K. Hansen and P. Salamon, "Neural network ensembles," IEEE Trans. Pattern Anal. Machine Intell., vol. 12, no. 10, pp. 993-1001, 1990.
- [14] R.E. Schapire, "The strength of weak learnability," Machine Learning, vol. 5, no. 2, pp. 197-227, June 1990.
- [15] [http://www.ics.uci.edu/\\$\sim\\$mllearn/MLRepository.html](http://www.ics.uci.edu/$\sim$mllearn/MLRepository.html)