Homogeneous Clustering Based Ensemble Classifier

Vaibhav Jain[1], P. Sai Sathvik[2], K.Sindhu[3], M.Sreelatha[4]

Department of Computer Science and Engineering
R.V.R. & J.C. College of Engineering Guntur, A.P
[1]vaibhavjain891992@gmail.com,[2]sathvikrao939@gmail.com,[3]sindhualapati926@gmail.com,
[4]lathamoturi@rediffmail.com

Abstract

The data set in the real world has overlapping class patterns. Classifying datasets with overlapping patterns is difficult. To classify the dataset with overlapping patterns the dataset is partitioned based on the class label and clustering is applied on each partition. The clustered data is given as an input to the base classifiers and the output of the base classifier is the cluster confidence vector. The result of base classifiers is given as an input to the fusion classifier. Fusion classifier maps the cluster confidence vector to class confidence vector. The proposed approach is verified on standard datasets from UCI machine learning repository and accuracies are measured, and compared.

Keywords- Homogeneous Clustering, Ensemble Classifier, Base classifier

I. INTRODUCTION

Ensemble methods use multiple learning algorithms to obtain better predictive performance than any of the constituent learning algorithms. An ensemble of classifiers is a set of classifiers whose individual decisions are combined to classify new examples. In order to combine the decisions of the individual base classifiers either fusion classifier or algebraic fusion can be used.

Ensembling can be carried out in two steps:

1. Construct a set of classifiers from the training data

2. Predict the class label of previously unseen examples by aggregating the predictions made by multiple classifiers.

Ensembling is a supervised learning algorithm. Ensemble of classifiers is more accurate than the individual classifiers. A necessary and sufficient condition for an ensemble of classifiers to be accurate than any of its individual members is if the classifiers are accurate and diverse. Two classifiers are diverse if they make different errors on new data points.

The latest work on ensemble classifiers resulted in the proposal of cluster oriented ensemble classifier [11] in which homogeneous clustering is performed on the dataset. This helps in significantly increasing the accuracy of the ensemble classifier but the loophole in lies in an unwarranted assumption that the number of clusters to be made for data items belonging to each class is equal. This assumption is unjustified because there may be the case that the number of patterns in the data items belonging to the same class may be different. Thus, clustering each partition (data items belonging to the same class) of the dataset into equal number of clusters in unfair.

In this paper, clustering each partition of the dataset into unequal number of clusters may yield a better result is proved. Nevertheless, there may be the cases for a few datasets where the overall classification accuracy is best when the number of clusters is equal for each partition. It may also be the case that the accuracy is best when no clustering is performed. This is an indication that each partition has only one pattern and clustering each partition doesn't aid in enhancing the overall accuracy. The proposal in the paper also assumes neural network as fusion classifier. Different classifiers are used as fusion classifier and the results of accuracies are tabulated when each different classifier is used as fusion classifier.

The rest of the paper is organized as follows: section 2 presents the related works. The proposed homogeneous clustering based ensemble classifier technique is discussed in section 3 and the experimental setup is presented in section 4. Section 5 describes the experimental results. Finally, section 6 concludes the paper.

II. RELATED WORK

Different ensemble classifier generation approaches are 1) Bagging (Bootstrap Aggregating) 2) Random forests 3) Boosting

Bagging [1] is an ensemble method that creates individuals for its ensemble by training each classifier on a random redistribution of the training set. The phases in the Bagging are i) Training Phase ii) Classification Phase. The algorithm for Bagging is as follows:

Training Phase:

1. Initialize the parameters
   - D=ø, the ensemble.
   - L, the number of classifiers to train.
2. For k=1 to L
   - Take a bootstrap sample S_k from the original dataset Z.
   - Build a classifier D_k using S_k as the training set.
   - Add the classifier to the current ensemble, $D=D \cup D_k$
3. Return D.

Classification Phase:

4. Run $D_1, D_2,\ldots,D_L$ on the input x.
5. The class with the maximum number of votes is chosen as the label for x.

Bagging is suitable for small datasets but doesn't scale well for the large datasets. The other problems with bagging are loss of interpretability and more computational complexity. In order to improve the performance of bagging, a variation to the bagging "random forests [2]" technique is used. Other variants of bagging include ordered aggregation [3], adaptive generation and aggregation approach [4]. Random forests are combination of tree predictors. This approach creates k trees where each tree is independently generated based on random decisions.

Boosting [5, 6] encompasses a family of methods. Boosting reduces bias in supervised learning. Boosting also uses weighting. Boosting technique weights models according to their performance. Variants of boosting include boosting recombined weak classifiers [7], weighted instance selection [8], Learn++ [9], and its variant Learn++, NC [10]. For larger datasets we can gain the accuracy by combining random forests with boosting.

All the three methods Bagging, Random Forests and Boosting does not provide any method to improve the learning process of base classifiers. In order to classify the dataset with overlapping patterns and to improve the learning process of base classifiers B. Verma and A. Rehman [11] proposed Cluster Oriented Ensemble Classifier.

One limitation with the cluster oriented ensemble classifier is the unwarranted assumption in the homogeneous clustering based ensemble classifier that the number of clusters to be made to data items belonging to different classes is equal.

III. PROPOSED METHOD

A. Motivation

Since the decision boundaries in the real world data sets are not simple, the dataset is clustered and then fed to base classifiers so that decision boundaries can be learned easily. The number of clusters made is proportional to the number of patterns. When homogeneous clustering is employed, the number of clusters made on instances of a particular class is proportional to the number of patterns in the instances of that particular class. But the number of patterns in data items belonging to different classes need not be equal. So a new ensemble classification approach using homogeneous clustering with different number of clusters made on instances belonging to different classes is proposed.

B. The Proposed Approach of Ensemble Classifier Model

Different phases in the proposed framework are

- Homogeneous Clustering
- Base classifier Training
- Fusion Classifier Training
- Prediction

1) Homogeneous Clustering

In homogeneous clustering the patterns belonging to each class are partitioned separately. Initially the data set is partitioned into N partitions where N represents the number of classes in the given dataset. Now each partition consists of the pattern that belongs to same class. Then each partition is divided into M clusters using k-means clustering algorithm where M is the number of clusters specified by the user. The clusters in each partition may not be equal.

After clustering the dataset the class label in the given dataset is replaced by the cluster label. Now the last column in the dataset consists of the cluster label.

2) Base Classifier Training

The clustered data is given as an input to the Base classifier. The different Base classifiers used are K-NN classifier, Neural Network Classifier and SVM classifier. The clustered data is given as input to the each base classifier and each base classifier produces cluster confidence vector.

3) Fusion Classifier Training

The cluster confidence vectors produced by different classifiers are combined and the combined clustered matrix is given as an input to the Fusion Classifier. The fusion classifier maps the cluster confidence values to the class confidence values i.e. fusion classifier performs cluster to class mapping. As an output the fusion classifier produces class confidence vector. Neural Network is used as Fusion Classifier.

4) Prediction

When a test tuple is fed to each of the base classifiers, cluster confidence vectors are obtained. These cluster confidence vectors are combined and fed to fusion classifier to obtain the final class confidence vector.

IV. EXPERIMENTAL SETUP

A number of experiments on seven benchmark data sets from UCI machine learning repository have been conducted to verify the strength of the proposed approach. A summary of the data sets is presented in Table I. 10-fold cross validation is used for reporting the classification results for all the data sets.

TABLE I DATASETS

Dataset	**#instances**	**#attributes**	**#classes**
Thyroid	215	5	3
Iris	150	4	3
Ionosphere	351	34	2
Wine	178	13	3
Cancer	699	10	2
Liver	345	7	2
Sonar	208	60	2

K-means clustering algorithm is used for clustering the datasets. Three base classifiers are used: the *k* Nearest Neighbour (*k*NN) classifier, Neural Network (NN) classifier, and the Support Vector Machine (SVM) classifier. All these classifiers are used simultaneously as base classifiers and they are used one at a time as a fusion classifier to find the impact of each of the classifiers in fusion process separately.

The neural networks are trained using tan sigmoid activation functions for the neurons and Levenberg-Marquardt backpropagation method for learning of the weights. We have used the radial basis kernel for SVM and the libsvm library [13] in all the experiments. The parameter *k* in *k*-NN is adjusted to that value of k where maximum accuracy was achieved with training data. Similarly, other parameters (such as sigma in RBF kernel of SVM) are also adjusted. All the experiments were conducted on MATLAB R2010b.

V. EXPERIMENTAL RESULTS

The following results are discussed:

1. The impact of making different number of clusters on instances of different classes
2. The impact of using different fusion classifiers.

A. Impact of Making Different Number of Clusters on Instances of Different Classes

The number of clusters made on training instances belonging to one class is different than that made to training instances belonging to other classes. The resultant accuracies can be seen in the Figures 1,2 for different datasets. Observe that the results included in Figures 1and 2 are using neural network as fusion classifier.

	1	2	3	4	5
1	92.5926	95	94.8148	92.5926	93.3333
2	94.4444	92.9630	94.9074	93.8889	93.7963
3	93.0556	93.6111	92.0370	92.6852	89.5370
4	92.6852	94.8148	92.5000	93.2407	91.4815
5	91.8519	93.5185	90.6481	91.3889	92.4074

(a) Ionosphere

	1	2	3	4	5
1	96.1346	96.8489	95.5611	94.8509	95.7081
2	96.2816	96.1387	95.4203	94.7039	96.4203
3	96.2795	95.7081	93.8427	95.7081	96.1366
4	95.4265	96.2836	94.7081	94.9979	94.7039
5	95.2795	96.9917	94.9938	94.5652	95.5714

(b) Cancer

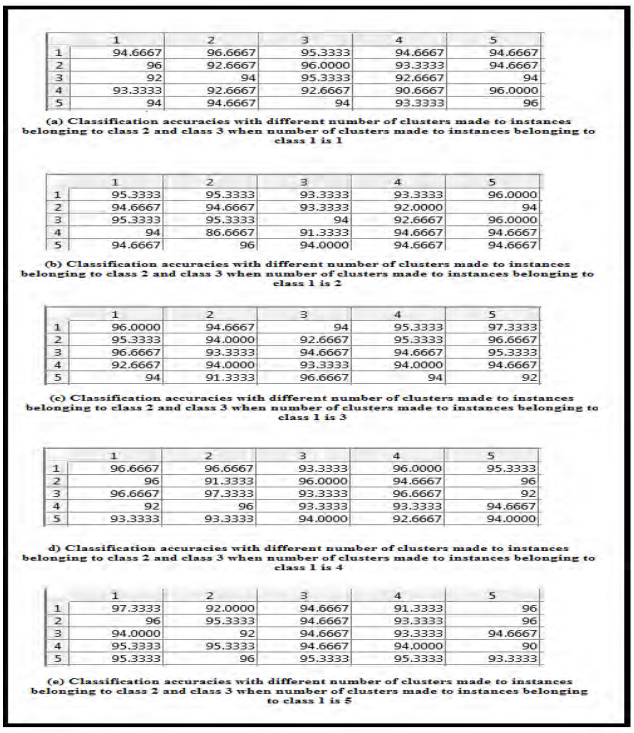
	1	2	3	4	5
1	70.8571	70.4286	68.1429	69.2857	65.4286
2	72.5238	69.8095	67	67.0952	66.0476
3	69.4286	66.8095	70.6667	66.8571	66.3810
4	69.2857	64.5714	68.5714	65	68.4286
5	70.7143	68.5238	66.2381	67.6667	63.4286

(c) Liver

	1	2	3	4	5
1	89.2732	87.5188	87.9950	86.9424	87.1429
2	87.8947	83.2832	84.7118	83.2832	85.4135
3	87.0927	86.1404	84.0351	86.6667	80.7018
4	84.6617	82.1303	87.8947	85.0877	82.0802
5	86.6165	86.5163	84.2356	83.3584	81.9048

(d) Sonar

Figure1. Classification accuracies of different datasets (with 2 classes)



	1	2	3	4	5
1	94.6667	96.6667	95.3333	94.6667	94.6667
2	96	92.6667	96.0000	93.3333	94.6667
3	92	94	95.3333	92.6667	94
4	93.3333	92.6667	92.6667	90.6667	96.0000
5	94	94.6667	94	93.3333	96

(a) Classification accuracies with different number of clusters made to instances belonging to class 2 and class 3 when number of clusters made to instances belonging to class 1 is 1

	1	2	3	4	5
1	95.3333	95.3333	93.3333	93.3333	96.0000
2	94.6667	94.6667	93.3333	92.0000	94
3	95.3333	95.3333	94	92.6667	96.0000
4	94	86.6667	91.3333	94.6667	94.6667
5	94.6667	96	94.0000	94.6667	94.6667

(b) Classification accuracies with different number of clusters made to instances belonging to class 2 and class 3 when number of clusters made to instances belonging to class 1 is 2

	1	2	3	4	5
1	96.0000	94.6667	94	95.3333	97.3333
2	95.3333	94.0000	92.6667	95.3333	96.6667
3	96.6667	93.3333	94.6667	94.6667	95.3333
4	92.6667	94.0000	93.3333	94.0000	94.6667
5	94	91.3333	96.6667	94	92

(c) Classification accuracies with different number of clusters made to instances belonging to class 2 and class 3 when number of clusters made to instances belonging to class 1 is 3

	1	2	3	4	5
1	96.6667	96.6667	93.3333	96.0000	95.3333
2	96	91.3333	96.0000	94.6667	96
3	96.6667	97.3333	93.3333	96.6667	92
4	92	96	93.3333	93.3333	94.6667
5	93.3333	93.3333	94.0000	92.6667	94.0000

d) Classification accuracies with different number of clusters made to instances belonging to class 2 and class 3 when number of clusters made to instances belonging to class 1 is 4

	1	2	3	4	5
1	97.3333	92.0000	94.6667	91.3333	96
2	96	95.3333	94.6667	93.3333	96
3	94.0000	92	94.6667	93.3333	94.6667
4	95.3333	95.3333	94.6667	94.0000	90
5	95.3333	96	95.3333	95.3333	93.3333

(e) Classification accuracies with different number of clusters made to instances belonging to class 2 and class 3 when number of clusters made to instances belonging to class 1 is 5

Figure2. Classification accuracies of iris dataset (with 3 classes) using homogeneous clustering based ensemble classifier with different number of clusters for instances of different classes as input.

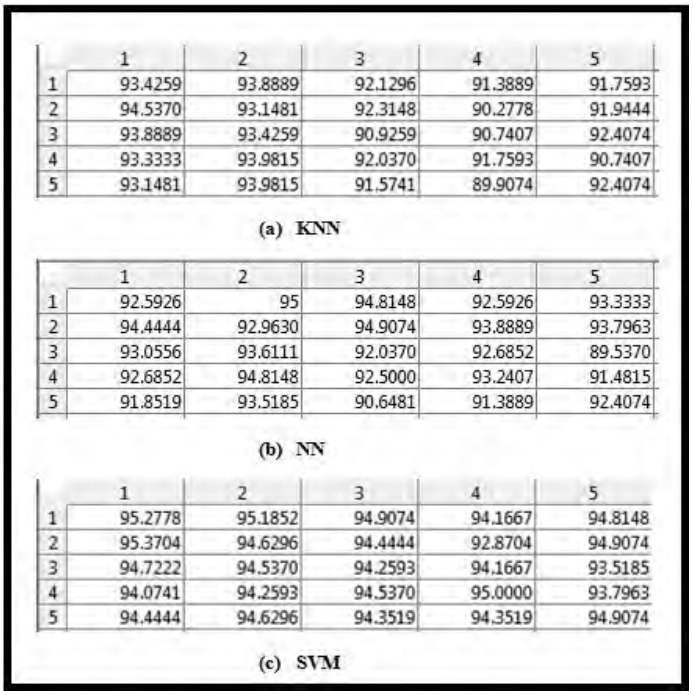
Table II presents comparison of maximum accuracies attained for different datasets by homogeneous clustering based ensemble classifier with instances of different classes clustered into equal and unequal number of clusters. The values in the brackets indicate the number of clusters made to instances belonging to a particular class. It can be observed the approach using unequal number of clusters performs better than that using equal number of clusters. On average, the approach using unequal number of clusters performs 1.23 percent better that the approach using equal number of clusters.

TABLE II. : Comparison of maximum accuracies attained for different datasets by Homogeneous clustering based ensemble classifier with equal and unequal number of clusters on instances of different classes

Dataset	**Homogeneous clustering with equal number of clusters**	**Homogeneous clustering with unequal number of clusters**
Iris	94.6667[1,1,1]	97.3333[3,1,5]
Wine	98.3333[1,1,1]	98.3333[1,1,1]
Thyroid	97.2727[2,2,2]	98.1818[2,3,3]
Ionosphere	93.2407[4,4]	95[1,2]
Liver	70.8571[1,1]	72.5238[2,1]
Cancer	96.1387[2,2]	96.9917[5,2]
Sonar	89.2732[1,1]	89.2732[1,1]

B. Impact of Using Different Fusion Classifiers

Figure 3 shows the classification accuracies of the proposed approach using different fusion classifiers for ionosphere dataset.



	1	2	3	4	5
1	93.4259	93.8889	92.1296	91.3889	91.7593
2	94.5370	93.1481	92.3148	90.2778	91.9444
3	93.8889	93.4259	90.9259	90.7407	92.4074
4	93.3333	93.9815	92.0370	91.7593	90.7407
5	93.1481	93.9815	91.5741	89.9074	92.4074

(a) KNN

	1	2	3	4	5
1	92.5926	95	94.8148	92.5926	93.3333
2	94.4444	92.9630	94.9074	93.8889	93.7963
3	93.0556	93.6111	92.0370	92.6852	89.5370
4	92.6852	94.8148	92.5000	93.2407	91.4815
5	91.8519	93.5185	90.6481	91.3889	92.4074

(b) NN

	1	2	3	4	5
1	95.2778	95.1852	94.9074	94.1667	94.8148
2	95.3704	94.6296	94.4444	92.8704	94.9074
3	94.7222	94.5370	94.2593	94.1667	93.5185
4	94.0741	94.2593	94.5370	95.0000	93.7963
5	94.4444	94.6296	94.3519	94.3519	94.9074

(c) SVM

Figure 3. Ionosphere dataset classification accuracies using different fusion classifiers

It can be observed from the figure that the maximum accuracy of 94.5370 using KNN is attained when number of clusters made to data items belonging to 1st class is 2 and number of clusters made to data items belonging to 1st class is 1. It is represented as 94.5370[2, 1]. Similarly, maximum accuracy using NN is 95.0000[1, 2], and that using SVM is 95.3704[2, 1].

Figure3 summarizes the results of ionosphere dataset and similar results of other datasets and presented in Tables III and IV. Table III shows the maximum accuracies attained using homogeneous clustering based ensemble classifier (at different number of clusters for instances of different classes) for different datasets and different fusion classifiers. It can be observed that the proposed approach using SVM as fusion classifier almost always gives better maximum accuracy than that using NN and kNN.

TABLE III. Maximum accuracies attained at different number of clusters per class using different fusion classifiers

Dataset	kNN	NN	SVM
Iris	98[5,4,4]	97.3333[3,1,5]	97.3333[1,1,5]
Wine	99.4444[1,2,3]	98.3333[1,1,1]	99.4444[2,1,5]
Thyroid	98.1818[2,5,1]	98.1818[2,3,3]	97.7273[1,4,2]
Ionosphere	94.5370[2,1]	95[1,2]	95.3704[2,1]
Liver	69.8095[1,3]	72.5238[2,1]	71.7619[1,3]
Cancer	96.5652[5,2]	96.9917[5,2]	97.4265[3,5]
Sonar	89.3233[2,1]	89.2732[1,1]	90.4762[3,4]

The Table IV shows the average accuracies attained for different datasets using ensemble classifier with homogeneously clustered data as input and different fusion classifiers. The average accuracy is the average of accuracies attained with different number of clusters per class. The table shows a clear dominance of using SVM as fusion classifier over other fusion classifiers. For all the considered datasets, the average accuracies (the average of accuracies obtained at different number of clusters) using SVM are the highest. On average, SVM fares 1.35 percent better than kNN and 1.51 percent better than NN.

TABLE IV. Average accuracy attained using different number of clusters per class while using different fusion classifiers

Dataset	kNN	NN	SVM
Iris	95.5253	94.3360	95.6907
Wine	96.6244	95.2494	97.3217
Thyroid	96.0732	95.0483	96.1538
Ionosphere	92.3630	92.9519	94.4852
Liver	64.7657	67.9676	68.7257
Cancer	95.5414	95.5706	96.6166
Sonar	86.5063	85.3033	86.9023

VI. CONCLUSION

A new homogeneous clustering based ensemble classifier is proposed and impact of using different fusion classifiers is discussed. The evidence from the experimental results shows that homogeneous clustering based ensemble classifier with unequal number of clusters on instances of different classes performs better than that with equal number of clusters by 1.23 percent. The proposed approach performs significantly better using SVM as fusion classifier. It performs 1.35 percent better than kNN and 1.51 percent better than NN as fusion classifier.

REFERENCES

[1] L. Breiman, Bagging predictors, Machine Learning, 24 (2)(1996), pp. 123–140
[2] L. Breiman, Random forests, Machine Learning, 45 (1) (2001), pp. 5–32
[3] G.M. Munoz, D.H. Lobato, and A. Suarez, “An Analysis of Ensemble Pruning Techniques Based on Ordered Aggregation,” IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 31, no. 2, pp. 245-259, Feb. 2009.
[4] L. Chen and M.S. Kamel, “A Generalized Adaptive Ensemble Generation and Aggregation Approach for Multiple Classifiers Systems,” Pattern Recognition, vol. 42, pp. 629-644, 2009.
[5] R.E. Schapire, The strength of weak learnability, Machine Learning, 5 (2) (1990), pp. 197–227
[6] Y. Freund and R.E. Schapire, “Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting,” J. Computer and System Sciences, vol. 55, no. 1, pp. 119-139, 1997.
[7] J.J. Rodriguez and J. Maudes, “Boosting Recombined Weak Classifiers,” Pattern Recognition Letters, vol. 29, pp. 1049-1059, 2008.
[8] N.G. Pedrajas, “Constructing Ensembles of Classifiers by Means of Weighted Instance Selection,” IEEE Trans. Neural Networks, vol. 20, no. 2, pp. 258-277, Feb. 2009.
[9] D. Parikh and R. Polikar, “Ensemble Based Incremental Learning Approach to Data Fusion,” IEEE Trans. Systems, Man, and Cybernetics, vol. 37, no. 2, pp. 437-450, Apr. 2007.
[10] M.D. Muhlbaier, A. Topalis, and R. Polikar, “Learn++.NC: Combining Ensemble of Classifiers with Dynamically Weighted Consult-and-Vote for Efficient Incremental Learning of New Classes,” IEEE Trans. Neural Networks, vol. 20, no. 1, pp. 152-168, Jan. 2009.
[11] Brijesh Verma, and Ashfaqur Rahman Cluster-Oriented Ensemble Classifier: Impact of Multicluster Characterization on Ensemble Classifier Learning IEEE transactions on knowledge and data engineering, vol. 24, no. 4, April 2012