A NEW FEATURE SET PARTITIONING METHOD FOR NEAREST MEAN CLASSIFIER ENSEMBLES

Abdullah[1], Ku Ruhana Ku-Mahamud[2], and Agung Sediyono[3]

[1]*Universitas Islam Indragiri, Indonesia, abdialam@yahoo.com*
[2]*Universiti Utara Malaysia, Malaysia, ruhana@uum.edu.my*
[3]*UniversitasTrisakti, Indonesia, trisakti_agung06@yahoo.com*

ABSTRACT. Nearest Mean Classifier (NMC) provides good performance for small sample size problem. However concatenate different features into a high dimensional feature vectors and process them using a single NMC generally does not give good results because of dimensionality problem. In this new method, the feature set is partitioned into disjoint feature subset based on diversity in ensemble. NMC ensemble is constructed by assigning each individual classifier in the ensemble with a cluster from different feature subset. The advantage of this method is that all available information in the training set is used. There is no irrelevant feature in the training set that was eliminated. Based on experimental results the new method shows a significant improvement with high statistical confidence.

Keywords: nearest mean classifier, feature set partitioning, ensemble classifier

INTRODUCTION

A well-known fast and simple classification algorithm is the Nearest Mean Classifier (NMC). NMC was introduced by Fukunaga (1990) as a classifier which uses the similarity between patterns to determine classification. For each class, NMC computes the class mean of the training patterns. Similarity values are obtained by calculating the Euclidean distance between the unknown patterns to the class mean of the training patterns. NMC classifies any unknown patterns to the class with the class mean closest to the test patterns. NMC has been successfully applied to many classification problems and showed good and robust performance (Shin & Kim, 2009). Furthermore NMC provides good performance for small sample size problem (Veenman & Tax, 2005). Small sample size problems are problems with the number of samples smaller than the number of features (Jain & Chandrasekaran, 1982).

Ensemble classifier aims to obtain the final classification decision by integrating the output of several individual classifiers (Han et al., 2007). The concept of ensemble classifier was first proposed by Suen et al. (1990) in order to improve the results of character recognition. In the literature, this research area is defined by a number of different names such as multiple classifier combination, multiple classifier system, combining classifiers, committees of learner, mixtures of experts, the consensus theory, hybrid methods, decision combination, multiple experts, cooperative agents, opinion pool and sensor fusion (Parvin et al., 2009). Regardless of the different names that have been defined, the ensemble classifier combines several classifiers to obtain the final classification result. Combining multiple classifiers is considered as a new direction for pattern classification. Ensemble classifier has been shown to be very helpful in improving the classification performance over single classifier approach (Du et al., 2009).

An approach that has been used to construct diverse classifier ensembles is the manipulation of input features. This approach assigns different subset of features among individual classifier in the ensemble and usually, the same base classifier is used. The main method of this approach is the random subspace method which assigns a random subset of the original features to individual classifier for the same training sample (Ho, 1998). However, feature subsets can overlap and their sizes are usually identical. Furthermore other methods that have similar idea with this method are the multiple feature subsets (Bay, 1998) and the attributes bagging (Bryll et al., 2003). These methods are similar in the way they assign features randomly to individual classifier in the ensemble. The differences are in the determination of subset and ensemble size. Another method that uses this approach is the feature set partitioning where the feature set is clustered into different feature subset. Classifier ensembles are constructed by assigning each individual classifier in the ensemble with a different feature subset from the pool of available features. The advantage of this approach is the used of available information in the training set. No irrelevant feature in the training set is eliminated. Irrelevant feature does not need to be eliminated in the combination of classifier because omitted feature may contain valuable information (Wang et al., 2005; Rokach, 2008). Therefore a new feature set partitioning method based on diversity measure for better NMC ensembles is proposed in this study.

PROPOSED METHOD

In this method, a group of classifier is built from the training set. A disjoint feature set decomposition is performed based on the original training set. Ensemble classifier is constructed based on the feature set decomposition. Prediction class label of unknown pattern is obtained by aggregating predictions using a combiner. In this study the normalized combination distance as has been adopted in Abdullah and Ku-Mahamud (2011) is used as a combiner. Figure 1 shows the framework of this method.

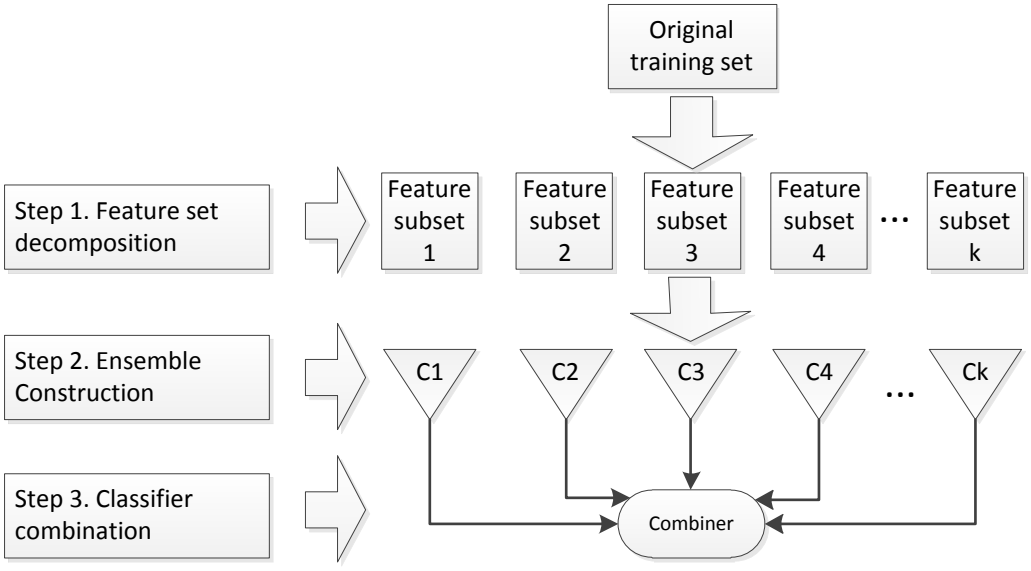

Figure 1. Framework of feature set partitioning method for classifier combination

An algorithm is developed to perform feature set decomposition based on diversity. Diversity is measured based on support diversity measure which is more frequently an agreement among individual NMC that provides small value diversity (Abdullah and Ku-Mahamud, 2012). Furthermore each NMC in ensemble is trained on a different projection of the original training set. The required inputs are the training set and class labels. The next step is to build

two feature subsets that give the maximum value of diversity to the ensemble. This step is repeated until all features are used for partition. Feature subsets that provide the maximum diversity measure is used to construct NMC ensembles. The flow chart for NMC ensembles construction is provided in Figure 2.

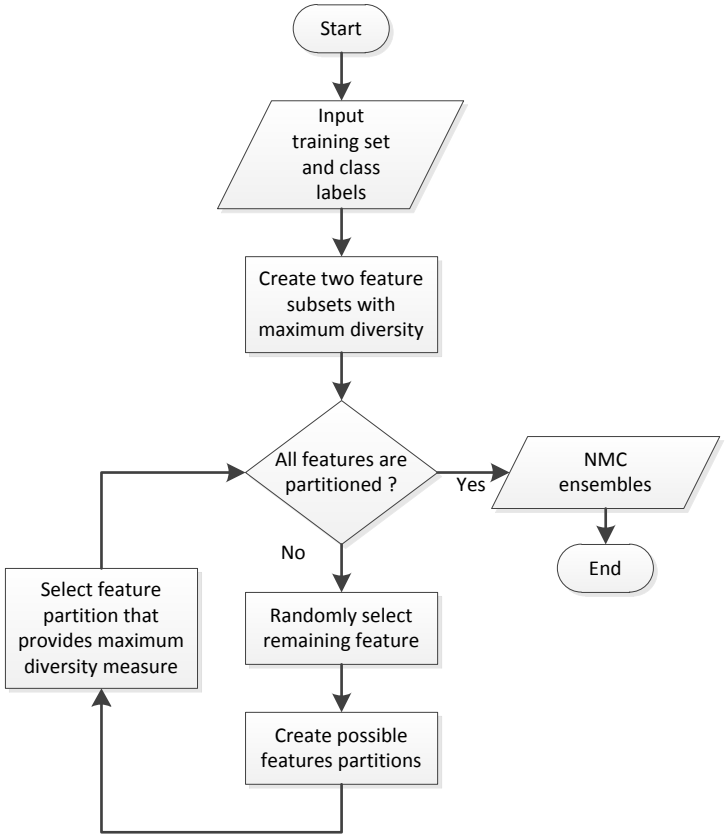

Figure 2. Flowchart of the diversity-based feature set partitioning algorithm for NMC ensembles

EXPERIMENTAL RESULTS

Classification experiments were performed on ten datasets. Pima, iris, wine, glass, liver, lenses, statlog (heart), ionosphere and soybean datasets were obtained from UCI machine learning repository, while fruit dataset is obtained by capturing images of fruit and are limited to several variants of apples, mangoes, oranges, pears and durian. Ten (10) experiments were performed on ten datasets to test the performance of the multiple NMC which has been constructed based on the new method and the results were compared with the performance of the original NMC. The average classification accuracy and standard deviation were computed. The results of the original nearest mean classifier accuracy on the ten datasets are presented in Table 1 while the results for multiple nearest mean classifiers (MNMC) accuracy using the diversity based feature partitioning algorithm on ten data sets are presented in Table 2.

Table 1. The accuracy of original nearest mean classifier (NMC)

Experiment #	Fruit	Pima	Iris	Wine	Glass	Liver	Lenses	Statlog	Ionosphere	Soybean
1	53.57	63.02	92.00	71.91	45.79	55.07	66.67	64.44	69.23	75.24
2	52.38	62.89	91.33	72.47	44.86	55.94	70.83	62.96	69.80	74.59
3	50.00	63.41	92.67	72.47	45.33	54.20	62.50	62.22	70.66	75.90
4	53.57	63.15	92.67	72.47	43.93	54.49	70.83	64.07	71.79	75.24
5	53.57	63.67	92.00	72.47	44.39	53.91	58.33	64.44	70.09	74.59
6	52.38	62.89	92.00	72.47	44.86	56.23	66.67	64.07	68.66	75.57
7	48.81	63.41	92.67	73.03	44.39	55.94	75.00	63.70	69.80	76.22
8	52.38	63.28	92.00	73.03	44.86	55.36	54.17	64.07	70.37	75.24
9	54.76	63.67	91.33	71.35	45.33	55.07	75.00	63.70	70.09	73.94
10	51.19	63.54	92.00	73.03	43.46	55.65	70.83	64.07	70.94	73.62
Average	**52.26**	**63.29**	**92.67**	**72.47**	**44.72**	**55.19**	**67.08**	**63.78**	**70.14**	**75.02**
Standard deviation	1.81	0.30	0.49	0.53	0.70	0.79	6.93	0.69	0.88	0.83

Table 2. The accuracy of multiple nearest mean classifiers combination (MNMC)

Experiment #	Fruit	Pima	Iris	Wine	Glass	Liver	Lenses	Statlog	Ionosphere	Soybean
1	95.24	67.06	87.33	95.51	37.85	53.33	75.00	79.26	80.06	75.57
2	96.43	67.71	86.67	93.26	48.60	55.07	62.50	82.59	76.92	76.55
3	97.62	68.49	88.00	93.82	47.20	52.75	75.00	84.44	76.35	75.90
4	92.86	68.23	87.33	92.70	47.66	54.49	75.00	84.81	73.79	72.64
5	96.43	67.84	86.00	92.13	48.60	49.57	87.50	85.93	78.35	75.57
6	86.90	67.71	87.33	95.51	50.00	55.36	70.83	82.22	79.49	74.59
7	94.05	67.45	86.00	93.26	50.00	53.04	58.33	82.22	75.50	71.34
8	97.62	67.32	86.67	92.70	48.60	56.23	66.67	82.22	79.20	81.11
9	97.62	68.10	86.67	94.38	47.66	53.91	54.17	85.19	74.07	73.29
10	96.43	67.58	87.33	93.82	50.00	54.49	66.67	82.59	79.49	76.55
Average	**95.12**	**67.75**	**86.93**	**93.71**	**47.62**	**53.83**	**69.17**	**83.15**	**77.32**	**75.31**
Standard deviation	3.29	0.43	0.64	1.15	3.58	1.85	9.66	1.96	2.33	2.68

The average accuracy of both methods is again presented in Figure 3.

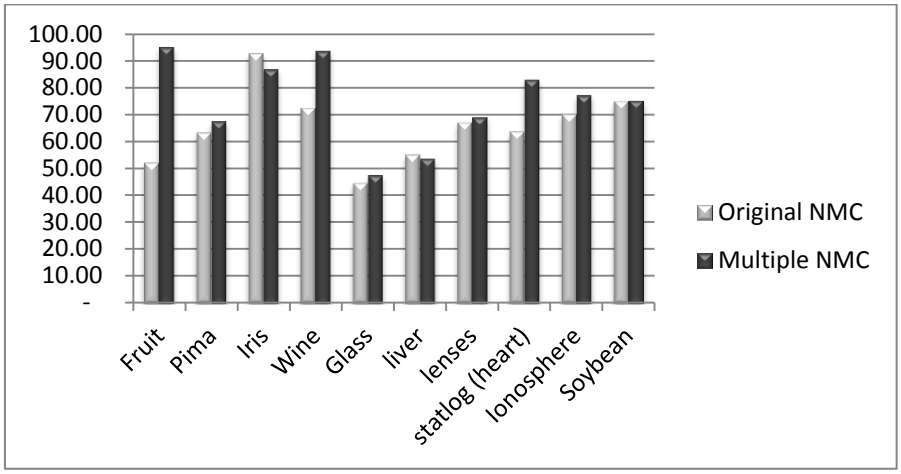

Figure 3. The bar chart of comparison of new multiple NMC with original NMC

Paired sample t-test is used to analysis accuracy improvement of NMC. Different treatment of paired samples t-test was performed i.e. before and after applying the algorithm to NMC. One-tail t-test was performed to see whether the average of the samples of MNMC is larger than the average of the sample of original NMC. Hypothesis for one-tail t-test for paired two samples are denoted as follows:

$H_o: \mu_1 = \mu_2$ (mean accuracies of original NMC and multiple NMC are the same)

$H_1: \mu_2 > \mu_1$ (mean accuracy of multiple NMC is greater than original NMC)

The hypothesis was tested statistically using a paired one-tail t-test, tested at the 5% significance level. The results of paired samples statistics and paired samples test are presented in Table 4 and Table 5 respectively.

Table 4. The output of paired samples statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	MNC	65.6620	10	13.42646	4.24582
	MNMC	74.9910	10	15.82364	5.00387

Table 5. The output of paired samples test

		Paired Samples Statistics							
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference		t	df	Sig.(2-tailed)
					Lower	Upper			
Pair 1	MNC-MNMC	-9.32900	14.57591	4.60931	-19.75598	1.09797	-2.024	9	.074

From the results, it can be seen that the sample mean for MNMC is 74.9910 while the sample mean of the original NMC is 65.6620 which is lower than the MNMC. The paired sample test shows that the two-tail probability value is 0.074 and thus the p-value = 0074/2 = 0.04 < 0.05 (5%). Therefore, H_o is rejected andH_1 is accepted. It can be concluded that the accuracy of NMC has significantly increased with 95% confidence after the implementation of the new feature set partitioning method.

CONCLUSION

A new feature set partitioning method for constructing NMC ensembles has been presented. The basic idea is to decompose the original features set into several subsets. Afterward every individual NMC in ensemble is trained on a different projection of the original training set, and then combine them. The method was evaluated on several datasets. The results show that implementation of this method to NMC significantly out performs original NMC. Results indicated that the proposed method can be used to create better NMC ensembles. Additional issue to be further studied is how the method can be implemented with other classifier.

REFERENCES

Abdullah & Ku-Mahamud, K.R. (2011). Combined nearest mean classifiers for multiple feature classification. *Proceedings of the 3rd International Conference on Computing and Informatics,* 8-13.

Abdullah & Ku-Mahamud, K.R. (2012). Prediction accuracy measurements for ensemble classifier. *Proceedings of the 6th Knowledge Management International Conference*, 102-107.

Bay, S.D. (1998). Combining nearest neighbor classifiers through multiple feature subsets. *Proceedings of the 17th International Conference on Machine Learning*, 37-45.

Bryll, R., Osuna, R.G. & Quek, F. (2003). Attribute bagging: improving accuracy of classifier ensembles by using random feature subsets. *Pattern Recognition*, 36, 1291-1302.

Du, P., Sun, H. & Zhang, W. (2009). Target identification from high resolution remote sensing image by combining multiple classifiers. *Proceedings of Multiple Classifier System*, 408-417.

Fukunaga, K. (1990). *Introduction to statistical pattern recognition* (2nd ed.).San Diego, CA, USA: Academic Press Professional Inc.

Han, D., Han, C. & Yang, Y. (2007). Multiple classifiers fusion based on weighted evidence combination. *Proceedings of the IEEE International Conference on Automation and Logistics*, 2138-2143.

Ho, T.K. (1998). The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(8), 832-844.

Jain, A.K. & Chandrasekaran, B. (1982). *Dimensionality and sample size consideration in pattern recognition practice*. Handbook of Statistics, 2, 835-855.

Parvin, H., Alizadeh, H. & Bidgoli, B.M. (2009).A new method for constructing classifier ensembles. *International Journal of Digital Content Technology and its Applications*, 3(2), 62-66.

Rokach L. (2008). Genetic algorithm-based feature set partitioning for classification problems. *Pattern Recognition*, 41(5), 1676-1700.

Shin, D. & Kim, S. (2009). Nearest mean classifier via-one class SVM. *Proceedings of the International Joint Conference on Computational Sciences and Optimization*, 1, 593-596.

Suen, C.Y., Nadal, C., Mai A., Legault R. & Lam, L. (1990). Recognition of totally unconstrained handwritten numerals based on the concept of multiple experts. *Proceedings of International Workshop on Frontiers in Handwriting Recognition*, 131-143.

Veenman, C.J. &Tax, D.M.J. (2005). A weighted nearest mean classifier for sparse subspaces. *Proceedings of Computer Vision and Pattern Recognition*, 2, 1171-1176.

Wang, L-J., Wang, X-L, & Chan, Q-C. (2005). GA-based feature subset clustering for combination of multiple nearest neighbors classifiers. *Proceedings of the 4th International Conference on Machine Learning and Cybernetics*, 2982-2987.