

The Implication of Data Diversity for a Classifier-free Ensemble Selection in Random Subspaces

Albert Hung-Ren Ko and Robert Sabourin

École de Technologie Supérieure, University of Quebec, Montreal, Quebec, H3C 1K3, Canada
albert@livia.etsmtl.ca, robert.sabourin@etsmtl.ca

Luiz E. Soares de Oliveira and Alceu de Souza Britto Jr.

Pontifical Catholic University of Parana, PR 80215-901, Curitiba, Brazil
alceu, soares@ppgia.pucpr.br

Abstract

Ensemble of Classifiers (EoC) has been shown effective in improving the performance of single classifiers by combining their outputs. By using diverse data subsets to train classifiers, the ensemble creation methods can create diverse classifiers for the EoC. In this work, we propose a scheme to measure the data diversity directly from random subspaces and we explore the possibility of using the data diversity directly to select the best data subsets for the construction of the EoC. The applicability is tested on NIST SD19 handwritten numerals.

1. Introduction

The goal of pattern recognition systems is to achieve the best possible classification performance. Since different classifiers usually make errors on different samples, we can combine classifiers to yield more accurate recognition rates. This approach is known as the Ensemble of Classifiers (EoC) method [3, 6]. Diverse classifiers can be created in several ways, such as Random Subspaces [2], Bagging and Boosting [5].

The two key issues that are crucial to the success of an EoC routine are the following: first, we need diversity for ensemble creation, because an EoC will not perform well without it [3, 5, 6, 8]; and second, we need to select classifiers once they have been created [5, 6, 8], because not all the classifiers created are useful. However, since not all the classifiers created will be used, time is spent in training classifiers that will not ultimately be used. Another is the evaluation

of high dimensional classifier combinations, since we need to evaluate different combinations of classifiers for ensemble selection after classifier training, and this evaluation will be very time-consuming in a large classifier pool. Hence, our question: Can we select data subsets for ensemble creation directly, instead of performing the ensemble creation/ensemble selection routine?

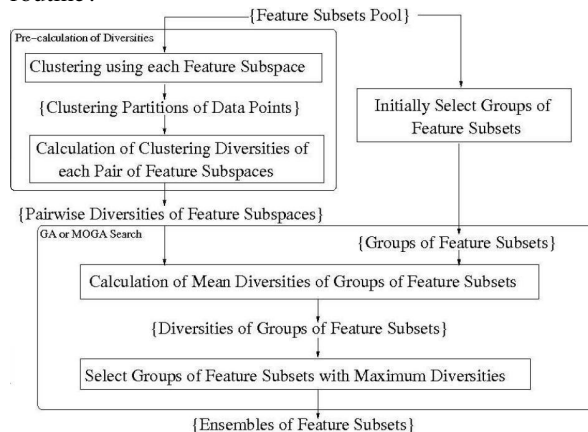


Figure 1. The proposed classifier-free ensemble selection scheme

We assume that data subset selection might be feasible through the evaluation of the data diversity of data subsets. We thus propose a data subset selection for the Random Subspaces ensemble generation method. Note that with this method data points might have relatively different distributions in the feature subspaces. This means that, by clustering these data points in different feature subspaces, we might have quite diverse clustering partitions. Since clustering diversities measure the diversity of these partitions,

they give an indirect indication of the data diversity of the feature subspaces. Fig. 1 shows the proposed classifier-free ensemble selection scheme, which is a feature subset selection in Random Subspaces. Note that the pre-calculation of diversities is carried out once for all, while GA or MOGA search are repeated from generation to generation.

Given a pool of feature subsets, we use a clustering algorithm with fixed parameters to form clusterings in feature subsets (Fig. 1). It is reasonable to assume that clustering diversity between different feature subsets also indicates their data diversity. This scheme will provide us with the following advantages:

1. By selecting the useful feature subsets, we can reduce the time needed for classifier training for ensemble creation.
2. By evaluating the pertinent feature subsets, we can significantly reduce the search space for ensemble selection.
3. Feature subset selection might be able to replace ensemble selection completely for Random Subspaces in some circumstances, and offers de facto classifier-free ensemble selection.

Our experimental results suggest that there is a strong correlation between classifier diversity and clustering diversity in Random Subspaces, and that clustering diversity does work for a classifier-free ensemble selection scheme. Here, we need to mention that the proposed strategy would not work for the Bagging and Boosting ensemble generation methods. Since Bagging and Boosting draw a certain proportion of the data points to train classifiers, it is quite possible that the distributions of data points are rather similar. Consequently, clustering these data points might not generate significantly different clustering partitions. More importantly, since Bagging uses various data points for each classifier, it is impossible for us to measure data diversity by clustering different parts of data points.

In the next section, we introduce general clustering diversity measures. In section 3, we investigate the possibility of ensemble selection using clustering diversity measures on NIST SD19 handwritten numeral digits. Discussion is provided in section 4 and our conclusion comprises the last section.

2. Clustering Diversity Measures

In general, given two clustering partitions C_i and C_k , we can apply clustering diversity to measure the diversity between them. Based on the pairwise counts, a number of clustering diversity measures are proposed [7]:

1) Wallace Indices

$$\text{Wallace-1}(W-1): W_i(C_i, C_k) = \frac{C_{11}}{C_{11} + C_{10}} \quad (1)$$

$$\text{Wallace-2}(W-2): W_k(C_i, C_k) = \frac{C_{11}}{C_{11} + C_{01}} \quad (2)$$

2) Fowlkes-Mallows Index (FM)

$$\begin{aligned} F(C_i, C_k) &= \frac{C_{11}}{((C_{11} + C_{10})(C_{11} + C_{01}))^{\frac{1}{2}}} \\ &= (W_i(C_i, C_k)W_k(C_i, C_k))^{\frac{1}{2}} \end{aligned} \quad (3)$$

3) Rand Index (RI)

$$R(C_i, C_k) = \frac{C_{11} + C_{00}}{\frac{C(C-1)}{2}} \quad (4)$$

4) Jacard Index (JI)

$$J(C_i, C_k) = \frac{C_{11}}{C_{11} + C_{01} + C_{10}} \quad (5)$$

5) Mirkin's Metric (MM)

$$K(C_i, C_k) = 2(C_{10} + C_{01}) = C(C-1)(1 - R(C_i, C_k)) \quad (6)$$

where: C_{11} is the number of data point pairs that are in the same cluster under both C_i and C_k ; C_{00} is the number of data point pairs that are in different clusters under both C_i and C_k ; C_{10} is the number of data point pairs that are in the same cluster under C_i , but not under C_k ; while C_{01} is the number of data point pairs that are in the same cluster under C_k , but not under C_i . Note that all these measures calculate the clustering diversity between two clusterings. In the case where there are more than two clusterings, the global clustering diversity is simply the mean of all clustering diversities between all clustering pairs. Given L clusterings, there are $\frac{L \times (L-1)}{2}$ clustering diversities $d_{12}, d_{13}, \dots, d_{(L-1)L}$ to be calculated, and the global clustering diversity \bar{d} will be its average:

$$\bar{d} = 2 \times \frac{\sum_{ij} d_{ij}}{L \times (L-1)}, \quad i \leq j \quad (7)$$

Now we want to check whether or not the clustering diversity of different feature subsets can be used as an objective function for classifier-free ensemble selection.

3. Evaluation of Objective Functions for Ensemble Selection

First, we need to evaluate the hypothesis that the clustering diversity of different feature subsets can be used as an objective function for ensemble selection in Random Subspaces. A clustering diversity was thus calculated based on the clusterings of these feature subsets, and served as an objective function for both the single genetic algorithm search (GA) and the multiobjective genetic algorithm search (MOGA). Once the feature subsets had been selected, we constructed corresponding classifiers using the selected feature subsets and evaluated the performance of the ensembles of these classifiers. At the same time, we also compare our classifier-free ensemble selection scheme with traditional classifier-based ensemble selection methods, which uses majority voting error (MVE) as the objective function for the GA and MOGA search algorithms.

The experiments were performed on a 10-class handwritten-numeral problem. The data were extracted from NISTSD19, essentially as in [9]. We first defined 100 feature subspaces for classifier-free ensemble selection (or feature subset selection), each feature subspace containing 32 features extracted from the total of 132 features. We used nearest neighbor classifiers ($K = 1$) for the KNN classifiers.

Four databases were used, including a training set with 5000 data points used to create 100 KNN in Random Subspaces, an optimization set with 10000 data points for the GA and the MOGA search, a validation set with 10000 data points to evaluate all the individuals according to the defined objective function, and then to store those individuals in a separate archive after each generation, and a test set with 60089 data points. For both the GA and MOGA search algorithms, we set at 128 the number of individuals in the population and 500 generations. The mutation rate was set to $\frac{1}{L}$, where L is the length of the mutated binary string [1], and the crossover probability was set to 50%. During the whole search, a threshold of 3 feature subsets or classifiers was applied as the minimum number of feature subsets or classifiers. All the experiments were carried out with 30 replications.

3.1 Single Genetic Algorithm for Ensemble Selection for Handwritten Numeral Recognition

For classifier-based ensemble selection, the EoCs selected by MVE achieved an average 96.45%

classification accuracy, while those selected by ME had only a 94.18% recognition rate (Table 1; Fig. 2). Note that the EoCs found by MVE have, in general, from 19 to 35 classifiers. However, for classifier free ensemble selection, the GA search led to the minimum number of feature subsets.

Table 1. The average recognition rates on test data of ensembles searched by GA. The simple majority voting was used as the fusion functions.

Classifier-based ensemble selection		
ME	MVE	
94.18 ± 0.00%	96.45 ± 0.05%	
Classifier-free ensemble selection		
W-1	W-2	FM
92.55 ± 0.55%	92.61 ± 0.43%	93.06 ± 0.14%
RI	JI	MM
92.25 ± 0.56%	92.22 ± 0.10%	93.03 ± 0.50%

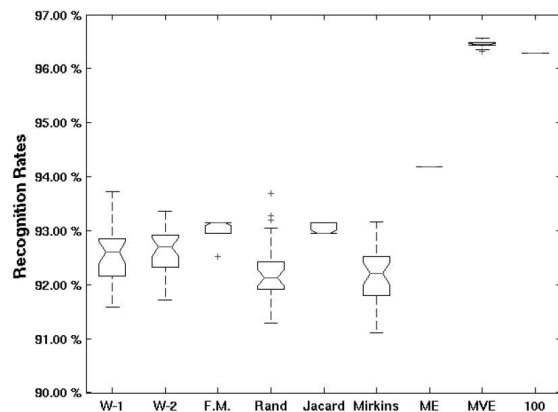


Figure 2. The average recognition rates achieved by EoCs selected by modified clustering diversities with single GA, compared with Mean Classifier Error (ME), Majority Voting Error (MVE), and the ensemble of all (100) KNN classifiers.

3.2 Multi-Objective Genetic Algorithms for Ensemble Selection for Handwritten Numeral Recognition

For classifier-free ensemble selection, the use of the MOGA search emphasizes the optimization of the clustering indices, as well as the maximization of the number of feature subsets. While the latter is no less relevant to better ensemble performance, it does avoid the problem of minimum ensemble size convergence that occurred in the GA search. While a MOGA search might not be necessary for classifier-based ensemble selection, we performed one nonetheless, so that we could compare the results of classifier-based ensemble selection with those of classifier-free ensemble selection.

First, we note that, because we used a MOGA, classifier-free ensemble selection with clustering diversity indices no longer converged to 3 feature subsets. This could allow further, more refined ensemble selection.

Moreover, we note that, in general, the feature subsets selected by classifier-free ensemble selection with clustering diversity indices construct adequate ensembles. The recognition rates achieved by these ensembles are very close to those achieved when all the classifiers are used (Fig. 3).

Table 2. The average recognition rates on test data of ensembles searched by MOGA. The simple majority voting was used as the fusion functions.

Classifier-based ensemble selection		
ME	MVE	
96.26 ± 0.08%	96.25 ± 0.04%	
Classifier-free ensemble selection		
W-1	W-2	FM
96.24 ± 0.08%	96.25 ± 0.06%	96.25 ± 0.08%
RI	JI	MM
96.23 ± 0.08%	96.26 ± 0.06%	96.19 ± 0.08%

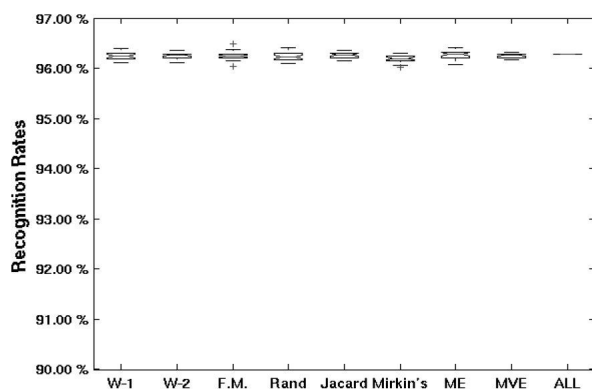


Figure 3. Box plot of the classifier-free ensemble selection schemes using MOGA compared with the classifier-based ensemble selection using Mean Error (ME) and Majority Voting Error (MVE) as objective functions

For classifier-based ensemble selection, ME also benefits from the MOGA scheme, and even slightly outperforms MVE as an objective function in a MOGA (See Table 2). By contrast, MVE did not perform quite as well as in a single GA, but the difference is rather small (0.20%). With a MOGA, MVE selected 49.25 classifiers on average, many more than it did with the simple GA.

The results of using the clustering diversities in classifier-free ensemble selection are encouraging, and all of them performed as well as the ensemble of all

classifiers, but the ensemble sizes were cut in half. Furthermore, there is no clear difference among the various clustering diversity measures (Fig. 3). This indicates that data diversity can be used to carry out ensemble selection in Random Subspaces, and that the proposed classifier-free ensemble selection scheme using clustering diversity measures as objective functions does work.

3.3 Classifier-Free Ensemble Selection Combined with Pairwise Fusion Functions for Handwritten Numeral Recognition

While MAJ is one of the fusion functions most often used for combining classifiers, it is not necessarily the optimum choice. If we apply other fusion functions - such as the pairwise fusion matrix with the majority voting rule (PFM-MAJ) [4] - the classifier-based ensemble selection using MVE might not be the best scheme. It turns out that the performances of ensembles selected by classifier-free ensemble selection can be further improved by using better fusion functions. As we can see in Table 3, the recognition rates of ensembles applying PFM-MAJ are apparently better than those applying the simple MAJ.

Table 3. The average recognition rates on test data of ensembles searched by MOGA. The pairwise confusion matrix applying the pairwise-majority voting was used as the fusion functions.

Classifier-based ensemble selection		
ME	MVE	
96.89 ± 0.05%	96.78 ± 0.09%	
Classifier-free ensemble selection		
W-1	W-2	FM
96.91 ± 0.05%	96.90 ± 0.04%	96.90 ± 0.04%
RI	JI	MM
96.90 ± 0.04%	96.89 ± 0.03%	96.88 ± 0.08%

Moreover, for the MOGA search, when PFM-MAJ was used as the fusion function, classifier-free ensemble selection using clustering diversity indices outperformed the classifier-based ensemble selection using MVE.

4 Discussion

In this work, we examined whether or not clustering diversity can represent the data diversity of different feature subsets in Random Subspaces, and whether or not the use of clustering diversity as the data diversity measure could allow us to apply a classifier-free ensemble selection scheme. First, for classifier-free ensemble selection, we used the single GA as the

search algorithm. We found that, with the clustering diversity indices as objective functions, it tends to converge to the minimum number of feature subsets, which makes a classifier-free ensemble selection scheme less useful.

Then, in order to compensate for the problem of the minimum feature subset convergence of the clustering diversities, we used the MOGA as the search algorithm. The clustering diversity measures yielded encouraging performances as objective functions for the classifier-free ensemble selection scheme.

However, we note that the proposed scheme for classifier free ensemble selection bears the additional cost of the clustering and on MOGA search. But, in general, the cost of the clusterings is much less than the cost of training classifiers such as the Support Vector Machine or the Multi-Layer Neural Network. Moreover, the comparison of the clusterings takes a relatively short time. For the MOGA search, the additional objective - the number of feature subsets - does not require complicated calculation.

The only major cost is the evaluation of the solutions found on the pareto front after the MOGA search. This requires the training of a classifier for each feature subset selected to evaluate the performances of ensembles, so that the best ensemble can be chosen. Compared with a traditional ensemble selection scheme, which requires the training of all classifiers and combinations of all the ensembles evaluated, the proposed scheme offers an interesting alternative. This approach will be especially attractive for tackling problems with a large classifier pool and time-consuming classifier training.

5 Conclusion

In this paper, we argue that clustering diversities actually represent the data diversities of different feature subsets in the Random Subspaces ensemble creation method. These data diversities can be measured with the help of clustering diversities without any classifier training. As a result, the feature subsets can be selected by clustering diversities to construct the classifiers in Random Subspaces.

Applying the MOGA search, we show that the ensembles selected by the clustering diversities had performances comparable to those selected by MVE, which is regarded as one of the best objective functions for ensemble selection [8]. The results are encouraging. Based on our exploratory work, we have drawn up some implications for the classifier-free ensemble selection approach:

- 1) In Random Subspaces, with the MOGA search the clustering diversity measures are good objective functions for ensemble selection.
- 2) In Random Subspaces, the ensembles selected by the different clustering diversity measures have so far been found to have similar performances based on the MOGA search.

Even though the clustering diversities might only be able to represent data diversities in Random Subspaces, for Bagging, which only use a part of the samples, there is still no adequate measure for their data diversities. It will be of great interest to figure out how to measure the data diversities in Bagging. Finally, we have to mention that, due to its special ensemble generating mechanism, the scheme is not likely to be applicable in Boosting.

Acknowledgment: this work was supported in part by grant OGP0106456 to Robert Sabourin from the NSERC of Canada.

References

- [1] A. E. Eiben, R. Hinterding, and Z. Michalewicz, "Parameter control in evolutionary algorithms", In *IEEE Transactions on Evolutionary Computation*, vol.3, no. 2, pp. 124-141, 1998
- [2] T.K. Ho, "The random space method for constructing decision forests", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 8, pp. 832-844, 1998
- [3] J. Kittler, M. Hatef, R. Duin, and J. Matas, "On Combining Classifiers", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 3, pp. 226-239, 1998
- [4] A. H. R. Ko, R. Sabourin, A. Britto Jr, A., L. Oliveira, "Pairwise Fusion Matrix for Combining Classifiers", *Pattern Recognition*, vol. 40, pp. 2198-2210, 2007
- [5] L. I. Kuncheva, M. Skurichina, and R. P. W. Duin, "An Experimental Study on Diversity for Bagging and Boosting with Linear Classifiers", *International Journal of Information Fusion*, vol. 3, no. 2, pp. 245-258, 2002
- [6] L. I. Kuncheva and C. J. Whitaker, "Measures of Diversity in Classifier Ensembles and Their Relationship with the Ensemble Accuracy", *Machine Learning*, vol. 51, no. 2, pp. 181-207, 2003
- [7] M. Meila, "Comparing clusterings", Technical Report 418, UW Statistics Department, 2002
- [8] D. Ruta and B. Gabrys, "Classifier Selection for Majority Voting", *International Journal of Information Fusion*, pp. 63-81, 2005
- [9] G. Tremblay, R. Sabourin, and P. Maupin, "Optimizing Nearest Neighbour in Random Subspaces using a Multi-Objective Genetic Algorithm", In *Proceedings of the International Conference on Pattern Recognition (ICPR 2004)*, pp 208-211, 2004.