A New Dynamic Ensemble Selection Method for Numeral Recognition

Albert Hung-Ren Ko, Robert Sabourin, and Alceu de Souza Britto, Jr.

LIVIA, ETS, University of Quebec
1100 Notre-Dame West Street, Montreal, Quebec, H3C 1K3 Canada

PPGIA, Pontifical Catholic University of Parana
Rua Imaculada Conceicao, 1155, PR 80215-901, Curitiba, Brazil.

{albert@livia.etsmtl.ca, robert.sabourin@etsmtl.ca,
alceu@ppgia.pucpr.br}

Abstract. An ensemble of classifiers (EoC) has been shown to be effective in improving classifier performance. To optimize EoC, the ensemble selection is one of the most imporatant issues. Dynamic scheme urges the use of different ensembles for different samples, but it has been shown that dynamic selection does not give better performance than static selection. We propose a dynamic selection scheme which explores the property of the oracle concept. The result suggests that the proposed scheme is apparently better than the selection based on popular majority voting error.

Key words: Fusion Function, Combining Classifiers, Diversity, Confusion Matrix, Pattern Recognition, Majority Voting, Ensemble of Learning Machines.

1 INTRODUCTION

The purpose of pattern recognition systems is to achieve the best possible classification performance. A number of classifiers are tested in these systems, and the most appropriate one is chosen for the problem at hand. Different classifiers usually make different errors on different samples, which means that, by combining classifiers, we can arrive at an ensemble that makes more accurate decisions [1, 8, 11]. In order to have classifiers with different errors, it is advisable to create diverse classifiers. For this purpose, diverse classifiers are grouped together into what is known as an Ensemble of Classifiers (EoC). There are several methods for creating diverse classifiers, among them Random Subspaces [6], Bagging and Boosting [10]. The Random Subspaces method creates various classifiers by using different subsets of features to train them. Because problems are represented in different subspaces, different classifiers develop different borders for the classification. Bagging generates diverse classifiers by randomly selecting subsets of samples to train classifiers. Intuitively, based on different sample subsets, classifiers would exhibit different behaviors. Boosting uses parts of samples to train classifiers as well, but not randomly; difficult samples have a greater probability of being selected,

and easier samples have less chance of being used for training. With this mechanism, most created classifiers will focus on hard samples and can be more effective.

There are two levels of problems in optimizing the performance of an EoC. First, how are classifiers selected, given a pool of different classifiers, to construct the best ensemble? Second, given all the selected classifiers, what is the best rule for combining their outputs? These two problems are fundamentally different, and should be solved separately to reduce the complexity of optimization of EoCs; the former focuses on ensemble selection [1, 11, 12] and the latter on ensemble combination, i.e. the choice of fusion functions [8, 12, 13]. For ensemble selection, the problem can be considered in two steps: (a) find a pertinent objective function for selecting the classifiers; and (b) use a pertinent searching algorithm to apply this criterion. Obviously, a correct criterion is one of the most crucial elements in selecting pertinent classifiers [1, 11, 12]. It is considered that, in a good ensemble, each classifier is required to have different errors, so that they will be corrected by the opinions of the whole group [8, 10–12, 15]. This property is regarded as the diversity of an ensemble. Diversity is thus widely used as objective function to select ensembles, but since diversity is not itself a fusion function, other authors proposed to directly use fusion functions such as a simple majority voting error rule (MVE) for ensemble selection.

However, the use of all these objective functions for ensemble selection is meant to construct one ensemble for all the samples. Intuitively, this is not the best way to combine classifiers, because different samples might be fit to different EoCs. Dynamic scheme explores the use of different classifiers for different samples [2–5, 7, 16]. Based on different features or different decision regions of each sample, a classifier is selected and assigned to the sample, some popular methods are a priori selection, a posteriori selection, overall local accuracy and local class accuracy [2–4, 16]. In general, their performances are compared with oracle, which is defined as the proportion of test samples that are at least correctly classified by one classifier in EoC. Nevertheless, against all expectations, it has been shown that dynamic selection has a large performance gap from the oracle [2], and moreover, it does not necessarily give better performance than static selection [4].

We note that most of dynamic selection schemes use the concept of the classifier accuracy on a defined neighborhood or region, such as local accuracy a priori or local accuracy a posteriori schemes [2]. These classifier accuracies are usually calculated with the help of KNN, and the use of these accuracies aims to realize an optimal Bayesian decision, but it is still outperformed by some static ensemble selection rule, such as MVE. This indicates a dilemma in estimation of these local accuracies, because their distribution might be too complicated to be well estimated. Interestingly, dynamic selection is regarded as an alternative of EoC [2, 3, 16], and is supposed to select the best classifier instead of the best EoC for a given sample. But, in fact, dynamic selection and EoC are not mutually exclusive. We believe that dynamic selection can also explore the strength of EoC.

We also note that, the oracle is usually regarded as a possible upper bound for EoC performances, and as far as we know, there is no effort made to explore the property of the oracle for dynamic selection. We argue that the complicated local classifier accuracy estimation can be actually carried out by oracle on a validation data set, and a

simple KNN method can allow the test data set to obtain the approximated local classifier accuracy from the validation data set. Here are the key questions that need to be addressed:

1. Can the concept of oracle be useful for dynamic selection?
2. Should we use the best classifier or the best EoC for dynamic selection?
3. Can dynamic selection outperform static selection?

To answer these questions, we propose a dynamic selection scheme which explores the property of the oracle concept, and compare the scheme with the static ensemble selection guided by different objective functions.

2 DYNAMIC CLASSIFIER SELECTION METHODS

2.1 Overall Local Accuracy (OLA)

The basic idea of this scheme is to estimate each individual classifier's accuracy in local regions of feature space surrounding a test sample, and then use the decision of the most locally accurate classifier [16]. The local accuracy is estimated as the percentage of training samples in the region that are correctly classified.

2.2 Local Class Accuracy (LCA)

This method is similar to the OLA, the only difference is that the local accuracy is estimated as the percentage of training samples with the respect to output classes [16]. In other words, we consider the percentage of the local training samples assigned to a class cl_i by this classifier that have been correctly labeled.

2.3 a priori selection method (a priori)

Instead of simply counting the percentage of training samples in the region that are correctly classified, we can calculate the average of probability outputs from correct classifiers. The probability can be further weighted by the distances between the training samples in the local region and the test sample. Consider the sample $x_j \in \omega_k$ as one of the k-nearest neighbors of the test pattern X, the $\hat{p}(\omega_k|x_j, c_i)$ provided by the classifier c_i can be regarded as a measure of the classifier accuracy for the test pattern X based on its neighbor x_j. Suppose we have N training samples in the neighborhood, then the best classifier C_* to classify the sample X can be selected by [2,4]:

$$C_* = \arg_i \max \frac{\sum_{j=1}^{N} \hat{p}(\omega_k|x_j \in \omega_k, c_i) W_j}{\sum_{j=1}^{N} W_j} \tag{1}$$

where $W_j = \frac{1}{d_j}$ is the distance between the test pattern X and the its neighbor sample x_j.

2.4 a posteriori selection method (a posteriori)

If the class assigned by the classifier c_i is known, $c_i(X) = \omega_k$, then this information can be exploited as well. Suppose we have N training samples in the neighborhood, and let us consider the sample $x_j \in \omega_k$ as one of the k-nearest neighbors of the test pattern X, then the best classifier $C_*(\omega_k)$ with the output class ω_k to classify the sample X can be selected by [2,4]:

$$C_*(\omega_k) = \arg_i \max \frac{\sum_{x_j \in \omega_k} \hat{p}(\omega_k|x_j, c_i)W_j}{\sum_{j=1}^{N} \hat{p}(\omega_k|x_j, c_i)W_j} \tag{2}$$

where $W_j = \frac{1}{d_j}$ is the distance between the test sample and the training sample.

3 K-Nearest-Oracles (KNORA) DYNAMIC CLASSIFIER SELECTION

All the above dynamic selection methods intend to find the most possibly correct classifier for a sample in a pre-defined neighborhood. But we propose another approach: Instead of finding the most suitable classifier, we select the most suitable ensemble for each sample.

The concept of the K-Nearest-Oracles (KNORA) is similar to those of OLA, LCA, a priori and a posteriori in terms of the consideration of the neighborhood of test patterns, but it distinguishes itself from the others by using directly the property of the oracle of the training samples in the region in order to find the best ensemble for a given sample. For any test data point, KNORA simply finds its nearest K neighbors in the validation set, figure out which classifiers correctly classify these neighbors in the validation set, and use them as the ensemble to classify the given pattern in the test set.

We propose four different schemes using KNORA:

1. KNORA-ELIMINATE (KN-E)
   Given K neighbors $x_j, 1 \leq j \leq K$ of a test pattern X, and suppose that a set of classifiers $C(j)$, $1 \leq j \leq K$ correctly classifies all its K nearest neighbors, then every classifier $c_i \in C(j)$ belonged to this correct classifier set $C(j)$ should gives a vote on the sample X. In case that none classifier can correctly classifies all K nearest neighbors of the test pattern, then we simply decrease the value of K until at least one classifier correctly classifies its neighbors.
2. KNORA-UNION (KN-U)
   Given K neighbors $x_j, 1 \leq j \leq K$ of a test pattern X, and suppose that the j nearest neighbor has been correctly classified by a set of classifiers $C(j)$, $1 \leq j \leq K$, then every classifier $c_i \in C(j)$ belonged to this correct classifier set $C(j)$ should gives a vote on the sample X. Note that since K nearest neighbors are considered, a classifier can have more than one vote if it correctly classifies more than one neighbor. The more neighbors that one classifier correctly classifies, the more votes this classifier will have for a test pattern.
3. KNORA-ELIMINATE-W (KN-E-W)
   The same as KNORA-ELIMINATE, but each vote is weighted by the distance between neighbor pattern x_j and test pattern X.

4. KNORA-UNION-W (KN-U-W)
   The same as KNORA-UNION, but each vote is weighted by the distance between neighbor pattern x_j and test pattern X.

4 EXPERIMENTS FOR DYNAMIC SELECTION ON HANDWRITTEN NUMERALS

4.1 Experimental Protocol for KNN

We carried out experiments on a 10-class handwritten numeral problem. The data were extracted from $NIST\ SD19$, essentially as in [14], based on the ensembles of KNNs generated by the Random Subspaces method. We used nearest neighbor classifiers ($K = 1$) for KNN, each KNN classifier having a different feature subset of 32 features extracted from the total of 132 features.

To evaluate the static ensemble selection and the dynamic ensemble selection, four databases were used: the training set with 5000 samples ($hsf\_\{0-3\}$) to create 100 KNN in Random Subspaces. The optimization set containing 10000 samples ($hsf\_\{0-3\}$) was used for genetic algorithm (GA) searching for static ensemble selection. To avoid overfitting during GA searching, the selection set containing 10000 samples ($hsf\_\{0-3\}$) was used to select the best solution from the current population according to the objective function defined, and then to store it in a separate archive after each generation. Using the best solution from this archive, the test set containing 60089 samples ($hsf\_\{7\}$) was used to evaluate the EoC accuracies.

We need to address the fact that the classifiers used were generated with feature subsets having only 32 features out of a total of 132. The weak classifiers can help us better observe the effects of EoCs. If a classifier uses all available features and all training samples, a much better performance can be observed [2, 3]. But, since this is not the objective of this paper, we focus on the improvement of EoCs by optimizing fusion functions on combining classifiers. The benchmark KNN classifier uses all 132 features, and so, with $K = 1$ we can have 93.34% recognition rates. The combination of all 100 KNN by simple MAJ gives 96.28% classification accuracy. The possible upper limit of classification accuracy (the oracle) is defined as the ratio of samples which are classified correctly by at least one classifier in a pool to all samples. The oracle is 99.95% for KNN.

4.2 Static Ensemble Selection with Classifier Performance

The majority voting error (MVE) was tested because of its reputation as one of the best objective functions in selecting classifiers for ensembles [12], it evaluates directly the global EoC performance by the majority voting (MAJ) rule. Based on this reason we tested the MAJ as the objective function for the ensemble selection. Furthermore, we tested the mean classifier error (ME) as well. The MAJ is also used as the fusion function.

In table 1 we observe that the MVE performs better than ME as an objective function for the static ensemble selection. The ensemble selected by MVE also outperforms that of all 100 KNNs.

Table 1. The recognition rates on test data of ensembles searched by GA with the Mean Classifier Error, Majority Voting Error.

Objective Functions	Min	Q_L	Median	Q_U	Max
Mean Classifier Error (ME)	94.18 %	94.18 %	94.18 %	94.18 %	94.18 %
Majority Voting Error (MVE)	96.32 %	96.41 %	96.45 %	96.49 %	96.57 %

It is clear that the MVE achieved the best performance as the objective function compared with traditional diversity measures. Given that the MAJ is used as the fusion function, this is not surprising.

4.3 Dynamic Ensemble Selection

Even though the MVE can so far find the best ensemble for the all samples, this does not mean that a single ensemble is the best solution for combining classifiers. In other words, each sample may have a different most suitable ensemble. It is our purpose to know whether the use of different ensembles on different samples can further increase the accuracy of the system.

Table 2. The best recognition rates of proposed dynamic ensemble selection methods within the neighborhood sizes $1 \leq k \leq 30$. RR= Recognition Rates.

	KN-E	KN-E-W	KN-U	KN-U-W
RR	97.52 %	97.52 %	97.25 %	97.25 %
K-value	7,8	7,8	1	1

Note that the dynamic ensemble selection does not use any search algorithm for the ensemble selection, because each sample has its own ensemble for the classifier combination. As a result, the repetition of the search was also not necessary.

For the dynamic selection, only three databases were used: the training set with 5000 samples ($hsf\_\{0-3\}$) to create 100 KNN in Random Subspaces, The optimization set containing 10000 samples ($hsf\_\{0-3\}$) was used for the dynamic ensemble selection, and the test set containing 60089 samples ($hsf\_\{7\}$) was used to evaluate the EoC accuracies. We tested our KNORA algorithm and compared it with other proposed schemes: the overall local accuracy (OLA), the local class accuracy (LCA), the local class accuracy a priori (a priori), and the local class accuracy a posteriori (a posteriori).

We note that most of the dynamic schemes are so far better than all tested objective functions for the static ensemble selection, except OLA and a priori methods. Both LCA and a posteriori schemes achieved very good performances, with 97.40% of the recognition rates. But the KNORA-ELIMINATE and KNORA-ELIMINATE-W have good performance as well, and with 97.52% it is the best dynamic selection scheme in our handwritten numeral problems (Table $2, 3$).

If we compare their performances in different neighborhood sizes, we can notice that while LCA and a posteriori dynamic selection schemes outperform the static GA

Table 3. The best recognition rates of each dynamic ensemble selection methods within the neighborhood sizes $1 \leq k \leq 30$ on Dynamic Ensemble Selection.

Methods	KNORA	OLA	LCA	a priori	a posteriori
Recognition Rates	97.52 %	94.11 %	97.40 %	94.12 %	97.40 %
K-value	7,8	30	1	30	1

selection with MVE as the objective function in a small neighborhood, their performances declined when the value k augments (Fig. 1). In this case, the static GA selection with MVE may still be better than LCA and a posteriori dynamic selection schemes. By contrast, KNORA-ELIMINATE has a more stable performance even when the value of k increases. It gives a better recognition rates than all other schemes on our experimental study, except when $k = 1$. But still, the stable performance of KNORA-ELIMINATE suggests that the dynamic selection schemes are worth for more attention.

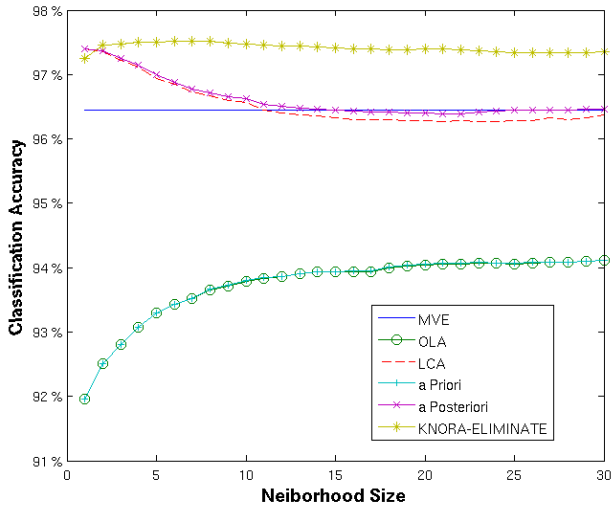

Fig. 1. The performances of various ensemble selection schemes based on different neighborhood sizes $1 \leq k \leq 30$ on NIST SD19 database. In the figure OLA overlaps with a priori selection.

5 DISCUSSION

In this paper, we propose a new dynamic ensemble selection scheme applying directly the concept of the oracle. Different from other dynamic selections, which use the es-

timated best classifier for a certain data point, the K-nearest oracle uses the estimated best EoCs for dynamic ensemble selection.

In our study of handwritten numeral digits, the proposed method apparently outperforms the static ensemble selection schemes such as the use of MVE or ME as the objective function in a GA search. Using the GA search, MVE can achieve 96.45% of the recognition rates, and ME attain can 94.18%. Nevertheless, with 97.52% of the recognition rates, KNORA-ELIMINATE is better then the evaluated static ensemble selection methods.

We note that OLA and a priori dynamic selection schemes were not as good as the static GA selection scheme with MVE. The OLA takes into account neither the class dependence, nor the weighting with the each classifier, and the a priori method ignores the class dependence. Since our experiment has high class dimension (10) and the ensemble pool size is quite large (100), it is not surprising that they do not perform well.

We also observe that KNIORA-UNION and KNORA-UNION-W are less performing than KNORA-ELIMINATE and KNORA-ELIMINATE-W. This might be due to the extreme elitism in the behavior of oracle.

Moreover, the KNORA-ELIMINATE also performs slightly better than other dynamic selection schemes. LCA and a posteriori schemes can achieve 97.40%, which is better than other static methods but inferior to the KNORA-ELIMINATE. However, the performance of the KNORA is still far from the oracle, which can achieve 99.95% of the recognition rates.

This might indicate that the behavior of the oracle is much more complex than a simple neighborhood approach can achieve, and it is not an easy task to figure out its behavior merely based on the pattern feature space.

6 CONCLUSION

We describe a methodology to dynamically select an ensemble for each data points. We find that by using directly the concept of the oracle, the proposed scheme has apparently better performances than the static ensemble selection schemes such as GA with MVE as the objective function. Moreover, the proposed schemes also perform slightly better than other dynamic selection methods in our study.

Besides this, the dynamic ensemble selection scheme has some additional advantages over the static ensemble selection schemes. For one, dynamic selection is pretty faster than some static selection - such as GA and exhaustive search. Also, the parameters embedded in the dynamic selection are much less than those of static selection. For example, considering the single GA search we need to adjust the mutation rate, the number of generation, the size of population size, and so on. All these make the optimization of the dynamic selection much easier.

Our study shows that a dynamic ensemble selection scheme can, in some cases, perform better than some static ensemble selection methods. Furthermore, our study suggests that an ensemble of classifier might be more stable than a single classifier in the case of a dynamic selection. Yet our method is limited by the uncertainty of the behavior of the oracle, since the attained recognition rates are still not close to that of

the oracle. We believe that this methodology can be much enhanced with theoretical studies on the connection between the feature subspaces and the classifier accuracies, the influence of geometrical and topological constraints on the oracle, better statistical studies to quantify the uncertainty of the oracle's behavior, and empirical studies in more real-world problems with various ensemble generation methods.

Acknowledgment

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

References

1. G. Brown, J. Wyatt, R. Harris and X. Yao, "Diversity Creation Methods: A Survey and Categorisation," *International Journal of Information Fusion*, vol. 6, no. 1, pp. 5-20, 2005
2. L. Didaci, G. Giacinto, F. Roli, G. L. Marcialis, "A study on the performances of dynamic classifier selection based on local accuracy estimation," *Pattern Recognition*, vol. 38, no. 11, pp. 2188-2191, 2005
3. L. Didaci, G. Giacinto, "Dynamic Classifier Selection by Adaptive k-Nearest-Neighbourhood Rule," *International Workshop on Multiple Classifier Systems (MCS 2004)*, pp. 174-183, 2004
4. G. Giacinto, F. Roli, "Methods for Dynamic Classifier Selection," *International Conference on Image Analysis and Processing (ICIAP 1999)* , pp. 659-664, 1999
5. T. Hastie, R. Tibshirani, "Discriminant Adaptive Nearest Neighbor Classification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*,vol. 18, no. 6, pp. 607-616, 1996
6. 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
7. T. K. Ho, J. J. Hull, and S. N. Srihari, "Decision combination in multiple classifier systems",*IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 16, no. 1, pp. 66- 75, 1994
8. 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
9. A. H. R. Ko, R. Sabourin, A. Britto Jr, "Combining Diversity and Classification Accuracy for Ensemble Selection in Random Subspaces" , *IEEE World Congress on Computational Intelligence (WCCI 2006) - International Joint Conference on Neural Networks (IJCNN 2006)*, 2006.
10. 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
11. 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
12. D. Ruta and B. Gabrys, "Classifier Selection for Majority Voting," *International Journal of Information Fusion*, pp. 63-81, 2005
13. D. M. J. Tax, M. Van Breukelen, R. P. W. Duin, and J. Kittler, "Combining Multiple Classifiers by Averaging or by Multiplying," *Pattern Recognition*, vol. 33, no. 9, pp.1475-1485, 2000

14. G. Tremblay, R. Sabourin, and P. Maupin, "Optimizing Nearest Neighbour in Random Subspaces using a Multi-Objective Genetic Algorithm," *In Proceedings of the 17th International Conference on Pattern Recognition (ICPR 2004)*, pp 208-211, 2004
15. D. Ruta and B. Gabrys, "Analysis of the Correlation between Majority Voting Error and the Diversity Measures in Multiple Classifier Systems," *In Proceedings of the 4th International Symposium on Soft Computing*, 2001
16. K. Woods, W. P. Kegelmeyer Jr, and K. Bowyer, "Combination of multiple classifiers using local accuracy estimates," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 4, pp. 405–410, 1997