Comparison of Classifier Selection Methods for Improving Committee Performance

Matti Aksela

Helsinki University of Technology, Neural Networks Research Centre
P.O.Box 5400, Fin-02015 HUT, Finland
matti.aksela@hut.fi

Abstract. Combining classifiers is an effective way of improving classification performance. In many situations it is possible to construct several classifiers with different characteristics. Selecting the member classifiers with the best individual performance can be shown to be suboptimal in several cases, and hence there exists a need to attempt to find effective member classifier selection methods. In this paper six selection criteria are discussed and evaluated in the setting of combining classifiers for isolated handwritten character recognition. A criterion focused on penalizing many classifiers making the same error, the exponential error count, is found to be able to produce the best selections.

1 Introduction

In an attempt to improve recognition performance it is a common approach to combine multiple classifiers in a committee formation. This is feasible if the outputs of several classifiers contain exclusive information. Often the focus of the research is on methods for combining the classifiers in the most effective manner, but it should not be forgotten that the committee's performance is highly dependent on the member classifiers used. In fact these two fundamental aspects in committee performance enhancement are often referred to as decision optimization and coverage optimization [1].

Instead of selecting member classifiers based solely on their accuracy, it may often be more effective to attempt to select the members based on their diversity, for which several measures have been presented [2–4]. Measuring the diversity of the member classifiers is by no means trivial, and there is a trade-off between diversity and member accuracy. Standard statistics do not take into account that for classification purposes a situation where identical correct answers are given differs greatly from the situation where identical erroneous answers are suggested, with the former being generally the best case and the latter the worst. For classification purposes it may be useful to examine especially the errors made.

Here six approaches to deciding on what subset of a larger set of member classifiers to use are examined. Three very different committee structures are briefly explained and used for evaluation with application to handwritten character recognition. Due to space constraints, readers are directed to the references for more thorough discussion on each member classifier and committee method.

2 Member Classifier Selection Criteria

Six different criteria for member classifier selection are presented here. The first three criteria are more traditional and have been gathered from literature. The latter three are novel and they have been designed based on the assumption of the significance of the classification errors being made. All except one of the presented approaches work in a pairwise fashion, where the result for a larger set is the mean of the pairwise measures for that set. The exception is the exponential error count in section 2.6, which compares all classifiers in the set simultaneously. As the minimum of the pairwise measures is always smaller or equal to the mean, the pairwise criteria are not suitable for selecting the size k of the classifier subset $C_1, \ldots, C_k$ from all of the K available classifiers. Hence it is assumed in all cases that the number of classifiers to be used is fixed in advance.

2.1 Correlation Between Errors

As it is reasonable to expect that the independence of occurring errors should be beneficial for classifier combining, the correlation of the errors for the member classifiers is a natural choice for comparing the subsets of classifiers. Here the correlation $\rho_{a,b}$ for the binary vectors v_e^a and v_e^b of error occurrence in classifiers a and b respectively is calculated as

$$\rho_{a,b} = \frac{\mathrm{Cov}[v_e^a, v_e^b]}{\sqrt{\mathrm{Var}[v_e^a]\mathrm{Var}[v_e^b]}}, \tag{1}$$

where Cov refers to covariance and Var variance. The best set is selected by choosing that with the minimal mean pairwise correlation.

2.2 *Q* Statistic

One statistic to assess the similarity of two classifiers is the Q statistic [2]. It is defined for two classifiers a, b as

$$Q_{a,b} = \frac{N^{11}N^{00} - N^{01}N^{10}}{N^{11}N^{00} + N^{01}N^{10}}\ , \tag{2}$$

where N^{11} is the number of times both classifiers are correct, N^{00} the number of times both classifiers are incorrect, and N^{10} and N^{01} the number of times when just the first or second classifier is correct, respectively. When the classifiers make just the same correct and incorrect decisions, it can be seen that the value of the Q statistic becomes one. Negative values indicate classifiers that make errors on different inputs. For sets of more than two classifiers the mean value of the pairwise Q statistics is considered to be the Q value for that set. The best subset of member classifiers is thus selected by minimizing the value of the Q statistic.

2.3 Mutual Information

As was suggested in [3], also a diversity measure based on calculating the mutual information of the classifiers results can establish a good set of member classifiers, as it by definition measures the amount of information shared between the classifiers. Hence minimizing the mutual information should produce a maximally diverse set of classifiers. The mutual information can be used for a measure of closeness and the pairwise mutual information between two classifiers a and b can be calculated as

$$I_{a,b} = \sum_{i=1}^{n} \sum_{j=1}^{n} p(c_i, c_j) \log \frac{p(c_i, c_j)}{p_a(c_i) p_b(c_j)} \ , \tag{3}$$

where n is the total number of classes and $c_i, i = 1, \ldots, n$ are the class labels.

In the experiments also the mutual information of the error occurrences has been calculated. There only two classes, correct or incorrect, are considered for each classifier. Both mutual information measures should be minimized to select the optimal subset of classifiers, again using the mean of the pairwise values for a larger set of classifiers.

2.4 Ratio Between Different and Same Errors

The worst possible setting for classifier combination is the situation where several classifiers agree on an incorrect result, and it is not nearly as fatal if they make errors to different labels. To explore this let us denote the count of how many times two classifiers made different errors at the same sample with $N^{00}_{\text{different}}$ and the count of how many times both classifiers made the same error with N^{00}_{same}. Now we can examine the ratio

$$r^{DSE}_{a,b} = \frac{N^{00}_{\text{different}}}{N^{00}_{\text{same}}} \ . \tag{4}$$

Again for more than two members the mean of the pairwise ratios is used. The best subset of classifiers can be selected through maximizing this ratio.

2.5 Weighted Count of Errors and Correct Results

One should consider taking into account information on also correct decisions in addition to the incorrect results, with more emphasis placed on the situation where classifiers agree on either the correct or incorrect result. One may simply count the occurrences of the situations and place suitable emphasis on the "both correct", a positive situation, and "both same incorrect", a negative situation:

$$r^{WCEC}_{a,b} = N^{11} + \frac{1}{2}(N^{01} + N^{10}) - N^{00}_{\text{different}} - 5N^{00}_{\text{same}} \ . \tag{5}$$

The weighting is arbitrary, and the presented values have been chosen as they are deemed suitable based on the reasoning to penalize errors, and especially same errors. For multiple classifiers, the mean of the pairwise counts is used. The optimal subset can be selected by maximizing the measure.

2.6 Exponential Error Count

As it is assumed that the member classifiers will hinder the classification the most when they agree on the same incorrect result, that situation can be given even more emphasis in the selection criterion. The errors can be counted and weighted by the number of classifiers making the error in an exponential fashion. The count of errors made by a total of i classifiers is denoted $N^0_{i\text{ same}}$ and added to the sum after rising to the ith power, or

$$r^{EXP}_{C_1,\ldots,C_k} = \frac{\sum_{i=1}^{k}(N^0_{i\text{ same}})^i}{N^1_{\text{all}}} \ . \tag{6}$$

This measure considers all member classifiers of the set at the same time, and the best combination is selected by minimizing the measure. Here also the correct classifications are taken into account by scaling the result with N^1_{all}, the number of samples for which every member classifier was correct.

It should be noted that more than one $N^0_{i\text{ same}}$ can be increased while processing one sample – if m classifiers agree on one erroneous result and n classifiers on another, both $N^0_{m\text{ same}}$ and $N^0_{n\text{ same}}$ are increased.

3 Committee Methods

Three quite different combination methods are used to evaluate the member classifier selection criteria. The plurality voting committee is a very simple method, while the Behavior-Knowledge Space (BKS) method [5] uses a separate training phase. As our experiments have focused on adaptive committee classifiers, also the run-time adaptive Dynamically Expanding Context (DEC) committee [6] is used for evaluation. For default decisions in the BKS and DEC methods, ie. for the situations where no rules yet exist, all member classifiers had been run on an evaluation set and ranked in the order of decreasing performance.

3.1 Plurality Voting Committee

The committee classifier simply uses the plurality voting rule to decide the output of the committee. This basic committee structure has been included because of its widespread use and familiar behavior.

3.2 BKS Committee

The Behavior-Knowledge Space (BKS) method [5] is based on a K-dimensional discrete space that is used to determine the class labels, with each dimension corresponding to the decision of one classifier. The result is obtained by first finding the focal unit in the K-dimensional space, the unit which is the intersection of the classifiers' decisions of the current input. Then if the unit has gathered samples and for some class c the ratio between the number of samples for class c and all gathered samples is above a threshold, class c is selected.

In the training phase the focal unit had collected the count of recognitions and counts for each true class. The output of the committee was the class with the highest probability in the focal unit, the one that had received most samples, as suggested in [5]. If the focal unit had not received any samples, the default rule of using the highest-ranking single classifier's result was used.

3.3 DEC Committee

The adaptive committee used is based on the Dynamically Expanding Context (DEC) algorithm [7]. The DEC principle had to be slightly modified to suit the setting of combining classifiers [6]. For this setting, a list of member classifiers' results is taken as a one-sided context for the first member classifier's result. The classifiers are used in the order of decreasing performance. To correct errors transformation rules consisting of a list of member classifier results as the inputs and the desired recognition result as the output are generated. Only rules whose output is included in the inputs may be produced.

Each time a character is input to the system, the existing rules are first searched through and the most specific applicable rule is used. If no applicable rule is found, the default decision is applied. For these experiments the default decision was taken to be a plurality voting decision among all member classifiers.

The classification result is compared to the correct class. If the recognition was incorrect, a new rule is created. The created rule always employs the minimal amount of context, ie. member classifier results, sufficient to distinguish it from existing rules. To make the rules distinguishable every new rule employs more contextual knowledge, if possible, than the rule causing its creation. Eventually the entire context available will be used and more precise rules can no longer be written. In such situations selection among multiple rules is performed via tracking correctness of the rules' usage.

4 Data and Member Classifiers

The data used in the experiments were isolated on-line characters in three separate databases. The preprocessing is covered in detail in [8]. Database 1 consists of 10403 characters written by 22 writers without any visual feedback. Databases 2 and 3 were collected with the pen trace shown on-screen and characters recognized on-line, with also the recognition results being shown. They contain 8046 and 8077 character samples, respectively, both written by eight different writers. All databases featured 68 character classes.

Database 1 was used solely for member classifier construction and training. Database 2 was used for training the BKS, the only committee method used requiring a separate training phase, and database 3 was used as a test set. The adaptive DEC committee performs run-time adaptation and creates writer-dependent rules during classification. The DEC committee was not trained beforehand in any way. For all member classifiers the sizes of the characters were scaled so that the the longer side of their bounding box was constant and the

Table 1. Member classifier performance

Classifier index	Member classifier	Error rate	Classifier rank
1	DTW PL Bounding box	23.06	4
2	DTW PL Mass center	20.02	2
3	DTW PP Bounding box	21.16	3
4	DTW PP Mass center	19.30	1
5	Point-sequence SVM	23.93	5
6	Grid SVM	26.49	6
7	Point-sequence NN	50.22	8
8	Grid NN	35.74	7

aspect ratio was kept unchanged. The accuracies of the individual member classifiers to be described below have been gathered into table 1.

4.1 DTW Member Classifiers

Four individual classifiers were based on stroke-by-stroke distances between the given character and the prototypes. Dynamic Time Warping (DTW) [9] was used to compute one of two distances, point-to-line (PL) or point-to-point (PP) [8]. In the PL distance the points of a stroke are matched to lines interpolated between the successive points of the opposite stroke. The PP distance uses the squared Euclidean distance between two data points as the cost function. The second variation was the use of either the 'Mass center' as the input sample's mass center or by 'Bounding box' as the center of the sample's bounding box, for defining the characters centers, thus creating four combinations. Database 1 was used for constructing the initial user-independent prototype set which consisted of 7 prototypes for each class.

4.2 SVM Member Classifiers

Two member classifiers based on Support Vector Machines (SVMs) were also included. The support vector machine classifiers were implemented using the libsvm version 2.36 SVM package [10]. The routines were slightly modified to accommodate the data used and to return a more information, but the classification and training routines were directly from the toolbox. Database 1 was used for training the SVM models.

The first of the SVM member classifiers takes its data as a list of points from the character. For this classifier, first the strokes of the characters are joined by appending all strokes to the first one. Then the one-stroked characters are transformed via interpolation or decimation to have a fixed number of points, for these experiments the point number was set to 30. Then the x and y coordinates of each point were concatenated to form a 60 dimensional vector of point coordinates, which were then used as data for the SVM classifier.

The second SVM member classifier takes a feature vector of values calculated from a grid representation of the character. For this classifier, a grid was formed and 17 values were calculated for each grid cell. These values include the sums of both negative and positive sin and cos of the slope of the line between the current and next point, the neighboring 8-neighborhood grid location the stroke moves from this location, the count of points in the cell and the count of pen-ups in the cell and character-wise means of these. A 3×3 grid was found to be the most promising of those tested (3×3, 5×5, 7×7, 10×10), and resulted in 153 dimensional data vectors.

4.3 NN Member Classifiers

Two member classifiers based on neural networks (NNs) were used. A fully connected feed-forward network structure was created using the Stuttgart Neural Network Simulator (SNNS) version 4.2 [11]. Database 1 was used for training.

The first NN classifier used the same preprocessing and feature vector type as the first SVM classifier, the coordinates of the one-stroke fixed-length characters were concatenated to form a 60 dimensional input vector. The number of output neurons was determined by the number of classes in the data, 68. A network using one hidden layer consisting of 100 neurons was used with 5000 epochs of training with the BackpropMomentum [11] learning algorithm.

The second NN used the grid-based approach with the same features as the second SVM classifier. Here a 5×5 grid was used resulting in 425 dimensional data. A network with two hidden layers of 100 neurons each and 68 output neurons was trained with the BackpropMomentum method for 1000 epochs.

5 Results

It can be seen in table 1 that the DTW-based classifiers are all better in individual performance than the other methods here. This stems from the fact that our own custom DTW classifier has been tuned for a prolonged period of time, while the other member classifiers were created from existing toolkits without nearly as much effort. Especially the performance of the point-sequence NN classifier is in itself unacceptable, but was included to examine the effect of a significantly worse but different member classifier.

The experiments were run using a fixed member classifiers set size of $k = 4$ member classifiers. The best combinations from all eight possible member classifiers produced by the selection criteria and the resulting accuracies from the three committee structures used for evaluation have been gathered into table 2. Also results using the four individually best member classifiers have been included as 'Best individual rates'. The three best results for each combination method using the brute force approach of going through all 70 possible combinations with each decision method have been collected into table 3.

In this case the correlation, Q statistic, and mutual information or errors measure selected exactly the same set of member classifiers, which is not surprising

Table 2. Comparison of selection criteria

Criterion	**Members**	**Vote**	**BKS**	**DEC**
Correlation	4,6,7,8	18.64	18.23	14.80
Q statistic	4,6,7,8	18.64	18.23	14.80
Mutual information	5,6,7,8	20.70	21.50	18.11
Mutual information of errors	4,6,7,8	18.64	18.23	14.80
Ratio between diff. and same errors	1,6,7,8	20.39	20.68	16.55
Weighted count of correct and err.	1,4,5,6	18.34	20.32	14.53
Exponential error count	4,5,6,8	16.45	18.13	14.12
Best individual rates	1,2,3,4	19.34	20.07	18.17

Table 3. Best brute force results

	Best		**Second best**		**Third best**	
Method	Members	Errors	Members	Errors	Members	Errors
Vote	4,5,6,8	16.45	2,5,6,8	16.66	3,5,6,8	17.14
BKS	2,3,4,7	17.46	2,3,6,8	17.83	3,4,6,8	17.85
DEC	4,5,6,8	14.12	2,5,6,8	14.44	4,5,6,7	14.49

considering their similar nature. The mutual information measure selected the worst performing classifiers. These criteria do not take into account the difference between errors and correct results – it is beneficial when members agree on correct results, but not when they agree on errors. All these criteria selected also the worst-performing point-sequence NN member classifier, which is not surprising considering its less similar, albeit due to numerous errors, results.

The approach of comparing the ratios between different and same errors does not perform well, providing the second-worst results. Also this criterion uses the clearly worst point-sequence NN classifier. The weighted count of errors and correct results criterion provides a combination that is second best for both the voting and DEC committees but notably poor for the BKS committee.

The exponential error count approach finds the best selection of all criteria in table 2. As can be seen in table 3, this criterion found the best combination of classifiers for this given task with respect to both the voting and DEC committees. This is in accordance with the initial assumption of the importance of the classifiers not making exactly the same mistakes too often.

An interesting difference of behavior can be noted with the BKS in comparison to the two other combination methods. The best member classifier set from table 3 for BKS is $\{2, 3, 4, 7\}$, the three members with the best individual accuracies and one with the worst. Presumably the different behavior is at least partly due to the fact that the separate training phase teaches the committee the types of errors that commonly occur. Hence the overall diversity of the set becomes a less important factor and the effects of the training weigh more on the final result when using a separate training phase. Also with the voting and

Table 4. Comparison of selection criteria without member classifier 7

Criterion	Members	Vote	BKS	DEC
Correlation	4,5,6,8	16.45	18.13	14.12
Q statistic	4,5,6,8	16.45	18.13	14.12
Mutual information	1,5,6,8	17.75	19.82	15.04
Mutual information of errors	4,5,6,8	16.45	18.13	14.12
Ratio between diff. and same errors	2,4,6,8	18.57	18.11	14.53
Weighted count of correct and err.	1,4,5,6	18.34	20.32	14.53
Exponential error count	4,5,6,8	16.45	18.13	14.12
Best individual rates	1,2,3,4	19.34	20.07	18.17

DEC approaches, the third-best combination was different, so clearly the optimal selection of member classifiers is also dependent on the combination method.

To evaluate the effect of just the one poorly performing, albeit diverse, classifier an additional experiment was run without using the point-sequence NN. The results are presented in table 4. It can be seen that the performance of the first five member selection criteria is greatly improved. Still the very simple ratio between different and same errors is clearly not sufficient for selecting a member classifier set here, when the results of all methods but the BKS are compared. But the correlation, Q statistic, and mutual information of errors criteria, who are still in agreement on the best selection of members, are now able to find the combination that produces the best results with the voting and DEC committees. Hence it may be a logical conclusion that these criteria are less robust with regard to member classifiers making a large number of errors.

6 Conclusions

Several member classifier selection criteria were examined, including statistical, information-theoretic and error counting measures. It appears that the more general criteria may be suboptimal for the specific case of classifier combining, especially when also poorly performing classifiers are included. When combining classifiers it would seem that the most important factor is that the classifiers as rarely as possible make exactly same mistakes, as these situations are the most difficult for the combining methods to anticipate. But it is also important that the member classifiers perform well. The best trade-off between accuracy and diversity was obtained with the suggested exponential error count criterion, which weighs identical errors made by the classifiers in an exponential fashion and normalizes the count with the number of cases where all members were correct. This method also showed robustness with respect to a very poorly performing member classifier.

One may ask what benefit there is in using a diversity measure instead of the decision mechanism directly when examining all subsets of classifiers. If it is possible to form a diversity measure for any subset from the pairwise diversity

measures of its members, noticeable computational benefits can be obtained. This is because going through all combinations of two classifiers results in significantly fewer possibilities than all combinations of a larger number of classifiers. For example in the presented experiments, with the subset size of 4 member classifiers from a total of 8 classifiers, the advantage is 28 vs. 70 combinations. The cost of forming the diversity measure of a subset from that of the pairs by averaging is insignificant. Naturally with very large numbers of member classifiers, a more evolved search scheme is necessary [4].

Still, if the objective is to optimize both the member classifier set and the decision method, the possibility of using a more general measure for the member classifier set selection should be considered. This may help refrain from excessive iteration of the two phases of optimization. But it must not be forgotten that the selection of member classifiers is dependent on the combination methods characteristics, a fact also concluded in [12] among others. A particular combination of classifiers, while optimal in some sense, does not guarantee the best results for all combination methods. However, a suitable measure may still provide some generalizational ability. Here the exponential error count has been shown to find a selection that consistently provides good results in the presented experiments.

References

1. Ho, T.K.: Multiple Classifier Combiation: Lessons and Next Steps. In: Hybrid Methods in Pattern Recognition. World Scientific Press (2002)
2. Kuncheva, L., Whittaker, C., Shipp, C., Duin, R.: Is independence good for combining classifiers. In: Proceedings of the 15th ICPR. Volume 2. (2000) 168–171
3. Kang, H., Lee, S.: An information-theoretic strategy for constructing multiple classifier systems. In: Proceedings of the 15th ICPR. Volume 2. (2000) 483–486
4. Roli, F., Giacinto, G.: Design of Multiple Classifier Systems. In: Hybrid Methods in Pattern Recognition. World Scientific Press (2002)
5. Huang, Y., Suen, C.: A method of combining multiple experts for the recognition of unconstrained handwritten numerals. IEEE Transactions on Pattern Analysis and Machine Intelligence **17** (1995) 90–94
6. Laaksonen, J., Aksela, M., Oja, E., Kangas, J.: Dynamically Expanding Context as committee adaptation method in on-line recognition of handwritten latin characters. In: Proceedings of ICDAR99. (1999) 796–799
7. Kohonen, T.: Dynamically expanding context. Journal of Intelligent Systems **1** (1987) 79–95
8. Vuori, V., Laaksonen, J., Oja, E., Kangas, J.: Experiments with adaptation strategies for a prototype-based recognition system of isolated handwritten characters. International Journal of Document Analysis and Recognition **3** (2001) 150–159
9. Sankoff, D., Kruskal, J.B.: Time warps, string edits, and macromolecules: the theory and practice of sequence comparison. Addison-Wesley (1983)
10. Chang, C.C., Lin, C.J.: Libsvm : a library for support vector machines version 2.36. http://www.csie.ntu.edu.tw/ cjlin/libsvm/ (2002)
11. Zell, A., Mache, N., et al, G.M.: Snns : Stuttgart neural network simulator. http://www-ra.informatik.uni-tuebingen.de/SNNS/ (2002)
12. Fumera, G., Roli, F.: Performance analysis and comparison of linear combiners for classifier fusion. In: Proceeding of S+SSPR2002. (2002) 424–432