Elimination and Combination of Classifiers in Multiple Classifier Systems

Te-Wei Chiang  
Department of Accounting  
Information Systems  
Chihlee Institute of Technology  
Taipei, Taiwan, R.O.C.  
ctw@mail.chihlee.edu.tw

Tienwei Tsai  
Department of Information Management  
Chihlee Institute of Technology  
Taipei, Taiwan, R.O.C.  
twt@mail.chihlee.edu.tw

Mann-Jung Hsiao  
Department of Information Management  
Kang-Ning Junior College of Medical Care and Management  
Taipei, Taiwan, R.O.C.  
hsiao@mis.knjc.edu.tw

Abstract
Traditional character recognition systems use a single classifier to determine the class of a given character. However, by using different types of classifiers simultaneously, the accuracy of classification could be improved. In this paper, we propose a classifier elimination approach based on correlation analysis and the derived heuristic rules to eliminate the redundant classifiers such that the succeeding decision combination process can be conducted in a more efficient and effective manner. Experimental results show our approach works well in the multiple classifier system.

Keywords
Multiple classifiers, classification, character recognition, decision combination, correlation analysis.

INTRODUCTION
In order to retrieve information from tremendous number of paper documents, transforming these documents into the contents that are accessible is inevitable. For over the last decade, optical character recognition (OCR) technique has been introduced as a practical approach for converting paper documents into computer codes.

Traditional character recognition systems use a single classifier to determine the true class of a given character. Recently, it has been observed that classifiers of different types complement one another in classification performance. Therefore, by using classifiers of different types simultaneously, classification accuracy could be improved. However, we have to coordinate the potentially conflicting decisions given by multiple classifiers. Ideally, the combination function should take advantage of the strengths of the individual classifiers, avoid their weaknesses, and improve classification accuracy.

Basically, we can apply the highest rank method to combine classifiers. The highest rank method is good for combining a small number of classifiers. Assume that for each input pattern $m$ classifiers are applied to rank a given set of classes. Thus each class received $m$ ranks. The minimum (highest) of these ranks is assigned to that class as its score. The classes are then sorted by these scores to derive a combined ranking for that input. One disadvantage of this method is that the combined ranking may have many ties.

Chiang et al. [2] proposed Limited Voting Method (LVM) and Weighted Voting Method (WVM). LVM was derived from majority vote. It is easy to implement and requires no training. However, it does not take into account the differences among individual classifier capabilities. In order to combine classifiers with inconsistent capabilities, LVM was modified by assigning weights to the rank scores produced by each classifier. This refined method is WVM. The weights in WVM are obtained by analyzing the recognition rate of each classifier in facing a certain class of characters, which reflects the relative significance of each classifier evaluated in the context of the combination.

T. K. Ho et al. [4] tried logistic regression, which has been widely used in the field of statistics, in order to solve alphanumeric character recognition problems. For each class, they used a binary variable $Y$ associated with to indicate an unknown testing sample belonging to the class ($Y=1$) or not ($Y=0$). In other words, for each class, the solution space has been partitioned into two subspaces (accepting region and rejecting region). However, this approach would encounter a serious problem as the number of classes becomes larger: the process of parameter estimation could not converge and therefore the parameters for logistic regression could not be found.

In addition to the issue of combining available classifiers, we have to consider how many classifiers are appropriate for the system with multiple classifiers. Since some classifiers may be highly dependent on other classifiers, the removal of redundant classifiers is also an important issue for efficient combination. In the domain of data mining, data reduction techniques [3] are applied to obtain a reduced representation of data set that is much smaller in volume, yet closely maintains the integrity of the original data. That is, mining on the reduced data set should be more efficient yet produce the same (or almost the same) analytical results. Therefore, some tech-
Symbols evolved from data mining, such as correlation analysis, is useful in dealing with the removal of redundant classifiers.

Based on above observations, we are motivated to develop an approach that can remove the redundant classifiers and combine the remained classifiers in an efficient and effective manner. The remainder of this paper is organized as follows. In Section 2, we define the symbols used in this paper. Section 3 introduces our approach to eliminate the redundant classifiers. Section 4 presents how to combine the decision of the reduced classifiers. Section 5 shows the experimental results of our system. Finally, Section 6 summarizes our approach.

**SYMBOL DEFINITION**

In this paper, \( e_i \) means classifier \( k \) where \( k = 1, ..., K \), and \( K \) is the total number of classifiers. \( C_1, ..., C_N \) are mutually exclusive and exhaustive sets of characters. \( N \) represents the total number of character classes. \( \Lambda = \{1, ..., N\} \) is a set which consists of all class index numbers. \( x \) denotes an input character and \( e_i(x) = \{r_i^1(x)\}, i = 1, 2, ..., N \) means that classifier \( k \) assigns the input character to each class \( i \) with a rank \( r_i(x) \). Then the problem becomes – When \( K \) classifiers give their individual decisions about the identity of an unknown input, how can these individual decisions be combined efficiently to produce a better decision? To formulate this problem, it becomes

\[
\begin{align*}
\{e_i(x) = r_i^1(x), ..., r_i^N(x) \} \\
\{e_2(x) = r_2^1(x), ..., r_2^N(x) \} \\
\vdots
\end{align*}
\]

\( e_i(x) = r_K^1(x), ..., r_K^N(x) \)

\( \rightarrow \)

\( R^1(x), ..., R^N(x) \)

\( \rightarrow \)

\( E(R^1(x), ..., R^N(x)) = j \)  \hspace{1cm} (1)

where \( R^i(x) \) is the combined rank of class \( i \) (\( 1 \leq i \leq N \)), and \( E \) is the decision making function of the multiple classifiers which gives \( x \) one definitive class \( j \) and \( j \in \Lambda \). The system is illustrated in Figure 1.

**ELIMINATION OF REDUNDANT CLASSIFIERS**

A classifier may be redundant if it can be “derived” from another one or its removal does not degrade the overall recognition rate.

**Correlation Analysis**

Some redundancies can be detected by correlation analysis. For example, given two variables, \( X \) and \( Y \), such analysis can measure how strongly one variable implies the other, based on the available data. Suppose that the data is represented by the pairs of values \( \{(x_i, y_i); i = 1, 2, ..., n\} \). Then the correlation between the two variables can be measured by the **sample correlation coefficient** [7], \( r \):

\[
r = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 \sum_{i=1}^{n} (y_i - \bar{y})^2}}, \hspace{1cm} (2)
\]

where \( \bar{x} \) and \( \bar{y} \) are the respective mean values of \( X \) and \( Y \). If the resulting value of Equation (2) is greater than 0, then \( X \) and \( Y \) are positively correlated, meaning that the values of \( X \) increase as the values of \( Y \) increase. The higher the value, the more each attribute implies the other. Hence, a high value may indicate that \( X \) or \( Y \) may be removed as a redundancy. Based on above observation, we can analysis the recognition results of each pair of individual classifiers during the training phase to find out those highly correlated classifiers. Therefore, the correlation between two classifiers \( e_p \) and \( e_q \), say \( r_{pq} \), can be found by

\[
r_{pq} = \frac{\sum_{i=1}^{n} (h_p^i - \bar{h}_p)(h_q^i - \bar{h}_q)}{\sqrt{\sum_{i=1}^{n} (h_p^i - \bar{h}_p)^2 \sum_{i=1}^{n} (h_q^i - \bar{h}_q)^2}}, \hspace{1cm} (3)
\]

where \( h_p^i \) and \( h_q^i \) are the respective hit values associated with classifier \( e_p \) and \( e_q \) in classifying the \( i \)-th training sample, and \( n \) is the number of training samples. \( h_p^i \) has the value 1 if the sample is correctly classified by classifier \( e_p \), and 0 otherwise.

Assume that the set of classifiers to be reduced is

---

**Figure 1. Decision combination of multiple classifiers**

- Input Character
- Reduced Candidate List \( S_0 \)
  - Classifier \#1
  - Classifier \#2
  - \( \ldots \)
  - Classifier \#N
- Sorted Candidate List \( S_1 \)
- Sorted Candidate List \( S_2 \)
  - \( \ldots \)
  - Sorted Candidate List \( S_N \)
- Decision Combination
- Identified Character

- \( e_i(x) = r_K^i(x), ..., r_K^N(x) \)
- \( \rightarrow \)
- \( R^i(x), ..., R^N(x) \)
- \( \rightarrow \)
- \( E(R^i(x), ..., R^N(x)) = j \)  \hspace{1cm} (1)
\( S_i = \{ e_1, e_2, ..., e_K \} \) and the set of correlation coefficients between each pair of classifiers is \( S_{ij} = \{ r_{ij} \} \), where \( i \leq j \leq K \). Since \( r_{ij} = r_{ji} \) and \( r_{ii} = 1 \), we only need to consider the \( r_{ij} \) in analyzing the correlations among all possible pairs of classifiers.

**Rules of Finding Redundant Classifiers**

From a series of experiments, we obtain four heuristic rules that are helpful in the process of finding redundant classifiers. They are summarized as follows:

- **R1:** The classifier with higher recognition rate is usually helpful and needs to be retained.
- **R2:** The classifier with lower recognition rate is usually redundant and needs to be removed.
- **R3:** If two classifiers are highly correlated, the one with lower recognition rate is usually redundant and needs to be removed.
- **R4:** If two classifiers are less correlated, these two classifiers may be the complements to each other.

Rules 1 and 2 are obtained intuitively. Rule 3 implies that if two classifiers have similar characteristics, each of which can be derived from the other and hence should be removed.

**Methods of Removing Redundant Classifiers**

Basically, to select useful classifiers or remove redundant classifiers, two basic heuristic methods can be applied:

1. **Stepwise forward selection:** The procedure starts with an empty set of classifiers. The best of the original classifiers is determined and added to the set. At each subsequent iteration or step, the best of the remaining original classifiers is added to the set.

2. **Stepwise backward elimination:** The procedure starts with the full set of classifiers. At each step, it removes the worst classifier remaining in the set.

The stopping criteria for the two methods may vary according to user’s requirement. The procedure may employ a threshold on the measure used to determine when to stop the classifier selection process. The measure could be the combined recognition rate or the number of selected classifiers.

Based on the above heuristic methods of classifier subset selection and the aforementioned heuristic rules, we develop the following methods that could remove redundant classifiers:

1. **FB:** At each step, forward select the classifier with the best recognition rate.
2. **FBL:** At each step, forward select two classifiers: the classifier with the best recognition rate and its least correlated one.
3. **FBM:** At each step, select the classifier with the best recognition rate and remove its most correlated classifier.
4. **BW:** At each step, backward remove the classifier with the worst recognition rate.

These methods are of great help in eliminating the redundant classifier in a multiple classifier system.

**DECISION COMBINATION**

To combine the set of reduced classifiers, two methods can be applied: Limited Voting Method (LVM) and Weighted Voting Method (WVM) [2].

**Limited Voting Method**

LVM derived from majority vote. The score obtained by class \( i \) \((1 \leq i \leq N)\) is given by

\[
M'(x) = \sum_{k=1}^{K} f(r_{ki}(x)),
\]

where \( f(r_{ki}(x)) \) is the rank score assigned to class \( i \) by classifier \( k \). The class with the highest rank obtains the largest score, say \( M_\text{max} \). Therefore, function \( f(r) \) can be defined as

\[
f(r) = \max(M_\text{max} - r + 1, 0).
\]

Finally, the decision rule applied can be defined as

\[
E(x) = \arg \max_{1 \leq i \leq N} \{ M'(x) \}.
\]

LVM is easy to implement and requires no training. However, it does not take into account the differences among individual classifier capabilities.

**Weighted Voting Method**

In order to combine classifiers with inconsistent capabilities, LVM is modified by assigning weights to the rank scores produced by each classifier. This refined method is called the Weighted Voting Method (WVM). Therefore, a modified function becomes

\[
M'(x) = \sum_{k=1}^{K} w_k f(r_{ki}(x)),
\]

where \( w_k \) is the weight (credibility) of classifier \( k \) in facing class \( i \). To obtain these weights, we gather
Seven classifiers were used in this experiment. Each classifier has different discrimination capability. We can find that individual classifiers, each of which was developed from different features of characters, can be ranked in the following order.

The correlation coefficients between any two classifiers are given in Table 2. From this table, we can conclude that the combination of two less-correlated classifiers is better than that of the loosely correlated combination pair, whose recognition rate is worse than that of the combination of all seven classifiers.

Table 1. The classification performance of seven individual classifiers

<table>
<thead>
<tr>
<th>Classifier</th>
<th>e_1</th>
<th>e_2</th>
<th>e_3</th>
<th>e_4</th>
<th>e_5</th>
<th>e_6</th>
<th>e_7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recognition rate</td>
<td>0.757</td>
<td>0.735</td>
<td>0.69</td>
<td>0.762</td>
<td>0.694</td>
<td>0.705</td>
<td>0.853</td>
</tr>
</tbody>
</table>

Table 2. Correlation coefficients between any two classifiers

<table>
<thead>
<tr>
<th>e_1</th>
<th>e_2</th>
<th>e_3</th>
<th>e_4</th>
<th>e_5</th>
<th>e_6</th>
<th>e_7</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.399</td>
<td>0.356</td>
<td>0.253</td>
<td>0.297</td>
<td>0.319</td>
<td>0.272</td>
</tr>
<tr>
<td>0.399</td>
<td>1</td>
<td>0.264</td>
<td>0.367</td>
<td>0.304</td>
<td>0.297</td>
<td>0.179</td>
</tr>
<tr>
<td>0.356</td>
<td>0.264</td>
<td>1</td>
<td>0.326</td>
<td>0.353</td>
<td>0.349</td>
<td>0.32</td>
</tr>
<tr>
<td>0.253</td>
<td>0.367</td>
<td>0.326</td>
<td>1</td>
<td>0.459</td>
<td>0.478</td>
<td>0.199</td>
</tr>
<tr>
<td>0.297</td>
<td>0.304</td>
<td>0.353</td>
<td>0.459</td>
<td>1</td>
<td>0.555</td>
<td>0.343</td>
</tr>
<tr>
<td>0.319</td>
<td>0.297</td>
<td>0.349</td>
<td>0.478</td>
<td>0.555</td>
<td>1</td>
<td>0.351</td>
</tr>
<tr>
<td>0.272</td>
<td>0.179</td>
<td>0.32</td>
<td>0.199</td>
<td>0.343</td>
<td>0.351</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 3. Recognition rates for the combination of any two classifiers

<table>
<thead>
<tr>
<th>e_1</th>
<th>e_2</th>
<th>e_3</th>
<th>e_4</th>
<th>e_5</th>
<th>e_6</th>
<th>e_7</th>
</tr>
</thead>
<tbody>
<tr>
<td>e_1</td>
<td>0.853</td>
<td>0.836</td>
<td>0.878</td>
<td>0.85</td>
<td>0.857</td>
<td>0.897</td>
</tr>
<tr>
<td>e_2</td>
<td>0.836</td>
<td>0.844</td>
<td>0.84</td>
<td>0.85</td>
<td>0.909</td>
<td></td>
</tr>
<tr>
<td>e_3</td>
<td>0.849</td>
<td>0.813</td>
<td>0.82</td>
<td>0.872</td>
<td></td>
<td></td>
</tr>
<tr>
<td>e_4</td>
<td>0.824</td>
<td>0.833</td>
<td>0.904</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>e_5</td>
<td>0.8</td>
<td>0.855</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>e_6</td>
<td>0.869</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

EXPERIMENTAL RESULTS

In our experiment, the total number of character classes is 3,235. Three sets of character images were used for training (set I, II) and testing (set III), each of which consists of 1,000 character images. Seven classifiers were used in this experiment.

Table 1 shows the classification result of individual classifiers, each of which was developed from different features of characters. We can find that each classifier has different discrimination capability.

The correlation coefficients between any two classifiers are given in Table 2. From this table, we can check whether two particular classifiers are highly correlated or not. For example, e_3 and e_6 are the most correlated classifiers, and e_2 and e_7 are the least correlated classifiers.

For the purpose of realizing the effectiveness of the combination of any two classifiers, the recognition rate after the combination of any two classifiers are given in Table 3. The decision combination strategy applied here is Weighted Voting Method (WVM). From Table 3 we can easily conclude that the classification performance after decision combination highly depends on the performance of each classifier which joins the combination. For example, if the best classifier e_7 is involved in a combination, then the recognition rate of this combination will be pretty good, no matter what the other classifier is. Also, the best recognition rate, 0.909, occurs under the combination of e_7 and e_2, both of which are good classifiers. This observation supports our proposed heuristic rule R1.

Further, the recognition rate of the combination pair (e_2, e_7) is better than that of pair (e_2, e_3) even though e_7 is better than e_2, which implies that the combination of two less-correlated classifiers would improve the effectiveness of the combination. Thus, heuristic rule R4 is reasonable. On the other hand, from Table 2 it is found that the most correlated classifiers are (e_3, e_6), whose recognition rate after combination is 0.8 (see Table 3), which is worse than that of the loosely correlated combination pair (e_2, e_7) even though e_6 is better than e_7. This evidence supports heuristic rule R3.

Table 4 shows the recognition rates combined from the classifiers after removing any two of them. From this table we can conclude that the removal of the worst classifiers would improve the recognition rate. For example, the best recognition rate, 0.937, occurs when the second worst classifier e_2 and the third worst classifier e_6 are removed at the same time. This observation supports our heuristic rule R2.

The combined recognition rates after eliminating the redundant classifiers, using our proposed heuristic methods, are given in Table 5. As to FB method, the recognition rate is almost the same as that of the original combination even though the number of classifiers is reduced to three. On applying methods FBL and BW, not only the number of classifiers has been reduced, but also the classification performance

<table>
<thead>
<tr>
<th>Method</th>
<th>ALL</th>
<th>FB</th>
<th>FBL</th>
<th>FBM</th>
<th>EW</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Classifiers</td>
<td>7</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Recognition Rate</td>
<td>0.937</td>
<td>0.935</td>
<td>0.939</td>
<td>0.933</td>
<td>0.939</td>
</tr>
</tbody>
</table>
has been improved. This experiment shows that there exists redundant classifiers and the elimination of those redundant classifiers would improve the classification performance.

CONCLUSIONS

This paper presents four heuristic rules that can help the elimination of redundant classifiers in a multiple expert decision combination system. Based on these rules, we also developed four methods for the removal of redundant classifiers. Experiment results show that the classification performance of the combination of the reduced set of classifiers provided by these methods is as good as or even better than that of the original set of classifiers.

REFERENCES


