Rank-Based Multiple Classifier Decision Combination: A Theoretical Study

Afşar Saranlı and Mübeccel Demirekler
Speech Processing Laboratory
Department of Electrical and Electronics Engineering
Middle East Technical University,
İsmet İnönü Bulv. 06531, Ankara, Türkiye.
{afsars, demirek}@metu.edu.tr

Abstract

This study presents a theoretical investigation of the rank-based multiple classifier decision problem for closed-set pattern identification. The problem of combining the decisions of more than one classifiers with raw outputs in the form of candidate class rankings is considered and formulated as a general discrete optimization problem with an objective function based on the total probability of correct decision. This formulation uses certain performance statistics about the joint behavior of the ensemble of classifiers, which need to be estimated from the cross-validation data. An initial approach leads to an integer (binary) programming problem with a simple and global optimum solution but of prohibitive dimensionality. Therefore, we present a partitioning formalism under which this dimensionality can be reduced by incorporating our prior knowledge about the problem domain and the structure of the training data. It is also shown that the formalism can effectively explain a number of successfully used combination approaches in the literature.

1. Introduction

The last decade has witnessed extensive research on the problem of combining the classification data supplied by a multitude of classifiers with the aim of improving the performance of the overall system. Contributions have been made or some form of classifier combination system have been attempted, among others within the fields of machine printed word/character recognition [15], handwritten character recognition [29, 16, 3, 23, 26, 18, 17], speaker recognition, [11, 6, 10, 22], face identification [1, 8], text to phoneme translation [28], remote sensing [4, 5], military target recognition [9] and biomedical image processing [19]. The neural networks community has also been active on this subject [27, 3, 21, 28, 2, 20, 13]. The diversity of the fields where encouraging results have been reported show that the methods of combining multiple classifiers is of considerable interest in many diverse applications of pattern recognition.

Xu and his colleagues have categorized multiple classifier combination systems with respect to the type of raw output information from each classifier, resulting in three such categories [29]. These are categories where the classifier outputs are single class labels (Type 1), rankings of a subset of candidate classes from highest to lowest "likelihood" (Type 2) and complete set of measurement values for the candidate classes leading to such rankings (Type 3). When a single classifier is considered, a final class label (the identified class) is obviously the only desired output. However, for combining multiple classifier outputs, only this abstract level may lead to a loss of valuable information. It should be advantageous to use classifier output forms with more information. Ho has shown [14] that rank-based combination is a good compromise which avoids output incompatibility and scaling problems while preserving valuable information about classifier behavior for imperfect classifiers in the raw classifier outputs. Despite the fact that there have been good theoretical attempts to analyze Type 1 and Type 3 systems, there have been few attempts to analyze rank-based combination systems. In [15, 14], Ho proposes, without attempting an in-depth theoretical analysis, to generalize the *Borda Count* method by linearly weighting the individual classifiers while Al-Ghoneim and Kumar proposes in [2], a method to train individual classifiers exploiting the knowledge that they will be involved in combination.

A good survey of existing rank-based combination methods is presented by Ho [15] where a number of new methods are also introduced. Among the two groups of methods discussed, the *class set reordering* methods are suitable for applications where a final unique class decision is required which is the case for our discussion. Methods belonging to this category are the *Highest Rank* method, the *Borda Count* method and the *Logistic Regression* method. The first two of these methods are simple methods which do not

use any classifier behavior observation while the last one introduced by Ho attempts to generalize the Borda Count method by linearly weighting the classifiers, incorporating the classifier behavior observation into the combination process.

The present study considers the rank-based multiple classifier decision problem and attempts a theoretical formulation treating this problem as one of discrete optimization. It will be shown that two of the rank-based combination methods discussed by Ho are special cases within this unifying formulation.

2. Problem Formulation and the Objective Function

We consider a closed-set pattern identification problem where patterns may come from P different classes $S_j, j = 1, 2, \cdots, P$ which we refer as the *candidate classes*. We assume there are Q classifiers $X_i, i = 1, 2, \cdots, Q$ involved in the classification process. Furthermore, x denotes a *pattern*, causing all classifiers to generate candidate class rankings. These rankings are transformed into a *rank score matrix* form $\mathbf{R}$ whose elements are positive integer *rank scores* r_{ji} with the highest score assigned to the highest ranking class. We define two random variables with values being indexes on an ordered set $\mathcal{S}$ of candidate classes: $\underline{s}_x$ denotes the true source class index, $\underline{d}$ denotes the final decision of the system. The pattern x processed by all classifiers results in a rank score matrix $\mathbf{R}$ which is the only input for final classification. Suppose that the ultimate objective function in classifier combination is to achieve the maximum rate of correct identification. Although other objective functions can also be defined, we will focus on this one for closed-set pattern recognition. The total probability of correct decision can be expressed as $P\{\underline{y} = 1\}$ where the random variable $\underline{y}$ is a binary valued *indicator* of the correct decision. The problem of finding the best rank-based multiple classifier decision process becomes one of maximizing $P\{\underline{y} = 1\}$ which is our objective function. This objective function is not very useful in this form and should be transformed in a form which contains free parameters for optimization as well as statistics about the classifier behavior. Expanding into a sum over source class and rank score matrix indexes and using the Bayes rule, the objective function becomes

$$P\{\underline{y} = 1\} = \sum_{j=1}^{P} \sum_{n=1}^{N} P\{\underline{d} = j | \underline{s}_x = j, \underline{r} = n\} P\{\underline{s}_x = j, \underline{r} = n\}. \quad (1)$$

Since the decision process to be found by definition uses only the rank score matrix, it is a deterministic function of $\underline{r}$. This means we have $P\{\underline{d} = j | \underline{s}_x = j, \underline{r} = n\} = P\{\underline{d} = j | \underline{r} = n\}$ which transforms (1) into

$$P\{\underline{y} = 1\} = \sum_{j=1}^{P} \sum_{n=1}^{N} P\{\underline{d} = j | \underline{r} = n\} P\{\underline{s}_x = j, \underline{r} = n\}. \quad (2)$$

In this form of the objective function, the first term $P\{\underline{d} = j | \underline{r} = n\}$ is directly linked with the decision process we are seeking. For a given deterministic decision process, these conditional terms are uniquely determined and are binary valued with values of "0" and "1". The second joint probability term $P\{\underline{s}_x = j, \underline{r} = n\}$ on the other hand is independent of the decision process and models the joint behavior of the ensemble of classifiers. Noting that it can also be expressed as $P\{\underline{r} = n | \underline{s}_x = j\} P\{\underline{s}_x = j\}$ it can be seen that this set of joint probabilities can be estimated if the classifiers are operated on cross-validation data with known class labels. Denoting the decision terms as our optimization variables b_{jn} and assuming that the *classifier observation statistics* have been properly estimated, the problem becomes

$$\max_{b_{jn}} \left\{ \sum_{j=1}^{P} \sum_{n=1}^{N} b_{jn} P\{\underline{s}_x = j, \underline{r} = n\} \right\}. \quad (3)$$

However, this is not an unconstrained optimization problem. Since the decision process can select only a single class label as the final decision for a given rank score matrix, our optimization variables are constrained by the set of constraints

$$\sum_{j=1}^{P} b_{jn} = 1 \quad \text{for} \quad n = 1, 2, \cdots, N. \quad (4)$$

3. Optimum Solution and the Curse of Dimensionality

The above constrained optimization problem seems to be a complex one. However, the set of constraints enable a simple global optimum solution. If the expansion in (3) is visualized as P lines and N columns with a term $b_{jn} P\{\underline{s}_x = j, \underline{r} = n\}$ at each row/column intersection, one can see that each column must have *only one* nonzero b_{jn} variable. Since all the estimated coefficients are non-negative, the global maximum can be achieved by selecting in each column the b_{jn} multiplying the maximum coefficient as "1" and all remaining variables as "0".

Once a set of values is determined for the b_{jn} variables, they correspond to a global optimum *decision process*. Consider an unknown pattern x processed by all the classifiers, leading to the rank score matrix $\mathbf{R}$. Once $\mathbf{R}$ is known, so is the index value n for $\underline{r}$. Among the P variables b_{jn} corresponding to this n value, the one being nonzero leads to the final classification by also determining the index value j for $\underline{s}_x$. The achieved global optimum decision may not be unique for the cases where at least one column in the aforementioned visualization contains more than one estimated coefficient which is maximum within the column. For these cases, the overall system is unable to discriminate among the indicated classes.

We have shown that an optimum solution is possible if we had the classifier observation statistics $P\{\underline{s}_x = j, \underline{r} = n\}$ estimated properly. This is the case for infinite data. Unfortunately, the number of such statistics is $P(P!)^Q$ which is prohibitively large even for small problems. Since such estimates should be done with limited data, a formalism of reducing the number of statistics to estimate is required. In the next sections, we will present such a formalism based on partitioning the observation space.

4. Partitioned Observation Space (POS) Approach

Consider the objective function in (3). The problem domain is composed of two main parts, the first one being the space spanned by the free variables b_{jn} and called the *Problem Parameter Space* while the second one being the space spanned by the estimated statistics $P\{\underline{s}_x = j, \underline{r} = n\}$ about the joint behavior of the classifiers, called the *Classifier Observation Space*. The statistics are called the *Classifier Observation Statistics*. For well behaving classifiers, the cross-validation samples tend to be clustered in the classifier observation space. Hence there will be no data for certain statistics while enough data will exist for some other statistics. A feasible idea is to partition the observation space such that generated partitions have enough cross-validation data for estimation. Such a partitioning may be done by incorporating our prior knowledge about the problem space or by using the actual distribution of the cross-validation data or in a hybrid manner. A formalism for exploiting these ideas follows.

We first define an *augmented event space* $\mathcal{F}$ composed of the compound events $(\underline{s}_x = j; \underline{r} = n)$. These are the most basic events, i.e., the *event atoms* in $\mathcal{F}$. Each such atom specifies the occurrence of the event "The source class for the pattern x was S_j and the set of classifiers generated the rank score matrix $\mathbf{R}_n$". This event space is finite with cardinality $P(P!)^Q$. Now assume that a *mapping* $\mathcal{W}$ partitions this event space into disjoint sets of event atoms. The name will denote both the partitioning and the mapping resulting from the partitioning. Assume that the partitioning $\mathcal{W}$ results in $M_{\mathcal{W}}$ partitions $W_1, W_2, \cdots, W_{M_{\mathcal{W}}}$ which are disjoint and with their union being the event space $\mathcal{F}$. The partitioning results in a new event space where the new basic events are the partitions. Hence $\mathcal{W}$ effectively defines a new random variable

$$\underline{g}_{\mathcal{W}} : \mathcal{S} \times \mathcal{R} \longmapsto \{1, 2, \cdots, M_{\mathcal{W}}\}. \tag{5}$$

whose values are indexes on an ordered set $G_{\mathcal{W}} = \{W_1, W_2, \cdots, W_{M_{\mathcal{W}}}\}$ and where $\mathcal{S}$ is the set of possible source classes while $\mathcal{R}$ is the set of possible rank score matrixes.

To incorporate the partitioning process into the objective function of (2) we observe that the random variable $\underline{g}_{\mathcal{W}}$ is a deterministic mapping from the values of $\underline{s}_x$ and $\underline{r}$. Therefore, the double sum can also be written by introducing the new random variable as

$$P\{\underline{y} = 1\} = \sum_{j=1}^{P} \sum_{n=1}^{N} P\{\underline{d} = j, \underline{s}_x = j, \underline{r} = n, \underline{g}_{\mathcal{W}} = \mathcal{W}(j, n)\} \tag{6}$$

which, by successively using the Bayes rule and the fact that the decision should be based on the rank score matrix only, may be put into the final form

$$P\{\underline{y} = 1\} = \sum_{j=1}^{P} \sum_{n=1}^{N} P\{\underline{d} = j | \underline{r} = n\} \cdot P\{\underline{s}_x = j, \underline{r} = n | \underline{g}_{\mathcal{W}} = \mathcal{W}(j, n)\} P\{\underline{g}_{\mathcal{W}} = \mathcal{W}(j, n)\}. \tag{7}$$

The first and last set of terms inside this expansion have the usual meanings of *decision variables* and *observation statistics* as in the previous expansion. However this time the observable events for modeling the joint classifier behavior in the observation space are the partitions W_m. This is a *coarser* resolution where the actual rank score matrixes are hidden behind observable partitions. The newly introduced set of terms in the middle of the expansion is a set of *transition terms* between the coarser resolution of the partitions and the finer resolution of the original event atoms. Clearly, the first terms are to be used as the optimization variables and the last terms will be estimated from the cross-validation data. Since a deliberate decision is made to keep the observation resolution at the coarser partition level, there is by definition no data left to determine the transition terms. In fact, due to this decision, one is *ignorant* about this finer detail. The transition terms allow us to formally introduce our ignorance within the Bayesian formalism, by assuming a uniform distribution within the partition, for the individual elements forming any such partition W_m. This takes the form

$$P\{\underline{s}_x = j, \underline{r} = n | \underline{g}_{\mathcal{W}} = m\} = \begin{cases} 0, & \text{if } (\underline{s}_x = j, \underline{r} = n) \notin W_m, \\ \frac{1}{|W_m|}, & \text{if } (\underline{s}_x = j, \underline{r} = n) \in W_m, \end{cases} \tag{8}$$

where $|W_m|$ is the cardinality of the partition W_m. Hence the only unknowns remaining within the expansion are the decision variables.

With this new expansion, a controlled tool to selectively decrease resolution on the observation and modeling of the classifier ensemble behavior is introduced. By the selection of the partitioning, it is possible to reduce the number of partitions, hence the events of the observation space by arbitrary orders of magnitude. Specifically, for the above expansion we have $M_{\mathcal{W}}$ statistics to estimate. For limited cross-validation data, a reduction in the number of statistics to estimate corresponds to an increase in the number of observations about each individual statistic, hence to an increase in the reliability of the estimate. It is well known in

pattern recognition literature that the reliability of the estimates is crucial to the generalization performance, hence to the identification performance of the system. Therefore, it is better to use a minimum number reliably estimated parameters instead of a larger number of unreliable ones [28, 12]. Although we have mentioned that the number of observable events can be arbitrarily reduced, this should be done by considering the amount of cross-validation data available. The new optimal solution based on observation statistics derived from a partitioning is sub-optimal as compared with the one based on the original statistics. Therefore, the nature of the partitioning is crucial for the usefulness of the resulting solution. The objective should be to maintain the maximum observation resolution *which is reasonable for the fixed amount of data available,* and not a finer one. It is also illogical to use a very coarse resolution while enough data for a finer one is available since, being unable to model the collective behavior of the set of classifiers properly will increase the deviation from the global optimum.

Clearly, the number of statistics to be estimated is reduced to $M_{\mathcal{W}}$, a number necessarily lower than the original cardinality. However, the double summation making up the objective function still has the original number of terms. Fortunately, even with this huge number of terms in the expansion, the optimum solution discussed previously may be converted into an algorithmic form so that only a small number of computations is necessary for making the optimum decision based on the estimated statistics. This algorithmic form can be summarized as follows: For each pattern, process it by all classifiers and generate the rank score matrix $\mathbf{R}$. Only for this specific $\mathbf{R}$, compute exactly P multiplier coefficients (the product of the last two terms in of the expansion in (7). I.e., one coefficient is computed for each candidate class. The final step is to decide on the class with the maximum coefficient. A total of at most P multiplications is involved. Note that the determination of the transition terms is only possible if the partitioning is based on a rule which can be easily applied when the rank score matrix is given.

5. A Sensible Partitioning: First Two Ranks

The partitioning used to decrease the number of statistics is often task dependent and a partitioning rule should be formulated for each application and even for each cross-validation data set. Often one may have prior insight into the task and classifiers involved before the observation statistics are collected. This may be incorporated into the solution by means of a partitioning of the observation space. The point will be illustrated with an example we call *first two ranks based partitioning*.

This partitioning is based on a general observation about the behavior of the classifiers. Assume it is intuitively expected that *the resolution below the topmost two ranks (largest two rank scores) is unreliable.* This is a reasonable expectation, for example for distance classifiers, since the separation between class models becomes less significant as the models become more and more separated from the pattern being classified in the feature space. Hence noisy features have greater chance to affect the lower rankings. Based on this expectation, we decide not to discriminate among the ranks lower than the second and group them to represent the *last rank*. The resulting new score assignment (hence partitioning) can be expressed as a simple rule given by

$$\hat{r}_{ij} = \begin{cases} r_{ij} - P + 3, & \text{if} \quad r_{ij} > P - 3, \\ 0, & \text{if} \quad r_{ij} \leq P - 3. \end{cases}$$

If an illustrative example of $P = 4$ classes and $Q = 2$ classifiers is considered, there are a total of $M_{\mathcal{W}} = 576$ partitions which are illustrated in Table 1. The first column is the set of event atoms inside a partition, the second column is a label for that partition and the last one is the partition random variable. Two shortcut notations are used for compactness of the example: Each row illustrates 4 partitions each corresponding to a particular subscript j of S_j and the corresponding value of $\underline{g}_{\mathcal{W}}$. The contents of each partition consisting of 4 rank score matrixes are illustrated by a *don't care* notation (i.e., elements denoted by the symbol $\times$ can take any allowable combination). Also the actual class names and rank score matrixes are used instead of the random variable names for clarity. There are a total of 576 resulting partitions (hence observation statistics) as compared to the original event space containing 2304 event atoms.

$$\left\{\left(S_{1,2,3,4},\begin{bmatrix}3&3\\2&2\\\times&\times\\\times&\times\end{bmatrix}\right)\right\}\mapsto\left(S_{1,2,3,4},\begin{bmatrix}2&2\\1&1\\0&0\\0&0\end{bmatrix}\right)\mapsto \underline{g}_{\mathcal{W}}=1,2,3,4$$

$$\left\{\left(S_{1,2,3,4},\begin{bmatrix}3&3\\2&\times\\\times&2\\\times&\times\end{bmatrix}\right)\right\}\mapsto\left(S_{1,2,3,4},\begin{bmatrix}2&2\\1&0\\0&1\\0&0\end{bmatrix}\right)\mapsto \underline{g}_{\mathcal{W}}=5,6,7,8$$

$$\vdots \qquad\qquad \vdots$$

$$\left\{\left(S_{1,2,3,4},\begin{bmatrix}\times&\times\\\times&\times\\2&2\\3&3\end{bmatrix}\right)\right\}\mapsto\left(S_{1,2,3,4},\begin{bmatrix}0&0\\0&0\\1&1\\2&2\end{bmatrix}\right)\mapsto \underline{g}_{\mathcal{W}}=573,574,575,576$$

Table 1. Illustration of First Two Ranks partitioning.

6. Special Cases as Specific Partitionings

In this section the theory will be linked with a number of existing rank-based multiple classifier decision techniques. It will be shown that they are special cases of the presented theory.

6.1. The Highest Rank Method

The Highest Rank method discussed in [15] is a simple technique of rank-based combination. Since it does not use any model of observed classifier behavior, it is not optimal in the general case. The procedure can be described verbally as "For each source class, select the highest of the rank scores assigned by all classifiers for that class as a new score. These new scores constitute a *max score* vector. The class with the maximum *max score* is selected." The status of this method with respect to the presented theory can be summarized as follows [24]:

Fact 1 *As a predefined decision process, the Highest Rank method coincides with a specific fixed set of values for the problem variables b_{jn}. The cases where the Highest Rank method is unable to make a decision because of more than one maximum in the sum score matrix correspond to more than one fixed sets of b_{jn} values.*

Fact 2 *The Highest Rank method does not make use of any observation on classifier ensemble behavior and hence is not in general optimum for the combination of non-ideal classifiers.*

The equivalence relation between the optimum solution resulting from the objective function expansion with respect to the partitioning W and the Highest Rank solution can be stated as a theorem whose proof is given elsewhere [24].

Theorem 1 *The Highest Rank method and the optimum solution to the classifier combination problem coincides in the context of maximizing the probability of correct decision expressed as in (7) only for a set of classifier observation statistics $P\{\underline{g}_{\mathcal{W}} = \mathcal{W}(j,n)\}$ which satisfy a fixed set of constraints.*

Table 2 illustrates by means of an example for $P = 3$ source classes and $Q = 2$ classifiers that the mapping generating the max score vector for the Highest Rank Method is a partitioning of the observation space. For this case, we let each max score vector to correspond to a partition. There are hence $M_{\mathcal{W}} = 42$ such partitions. A number of illustrative partitions are given in the Table. The same shortcut notation as in the First Two Ranks case except the don't care notation is again used here.

6.2. The Borda Count Method

This is yet another popular rank-based multiple classifier decision method. It is a slightly generalized majority voting technique from Group Decision Theory [15, 7]. Since it is simple to implement, it has been used as a popular rank-based technique. The Borda Count method uses the full rank score matrix $\mathbf{R}$ such that rank scores for each class are summed up across different classifiers. Therefore a *sum score* is obtained for each candidate class. The

$$
\begin{array}{ll}
\left\{\left(S_{1,2,3};\begin{bmatrix}2&2\\1&1\\0&0\end{bmatrix}\right)\right\}\mapsto & \left(S_{1,2,3};\begin{bmatrix}2\\1\\0\end{bmatrix}\right)\mapsto \underline{g}_{\mathcal{W}}=1,2,3\\
\left\{\left(S_{1,2,3};\begin{bmatrix}2&2\\1&0\\0&1\end{bmatrix}\right);\left(S_{1,2,3};\begin{bmatrix}2&2\\0&1\\1&0\end{bmatrix}\right)\right\}\mapsto & \left(S_{1,2,3};\begin{bmatrix}2\\1\\1\end{bmatrix}\right)\mapsto \underline{g}_{\mathcal{W}}=4,5,6\\
\vdots & \vdots\\
\left\{\left(S_{1,2,3};\begin{bmatrix}2&0\\1&0\\0&2\end{bmatrix}\right);\left(S_{1,2,3};\begin{bmatrix}2&0\\1&1\\0&2\end{bmatrix}\right);\left(S_{1,2,3};\begin{bmatrix}2&0\\0&1\\1&2\end{bmatrix}\right);\right. & \\
\left.\left(S_{1,2,3};\begin{bmatrix}1&2\\0&1\\2&0\end{bmatrix}\right);\left(S_{1,2,3};\begin{bmatrix}0&2\\1&1\\2&0\end{bmatrix}\right);\left(S_{1,2,3};\begin{bmatrix}0&2\\1&0\\2&1\end{bmatrix}\right)\right\}\mapsto & \left(S_{1,2,3};\begin{bmatrix}2\\1\\2\end{bmatrix}\right)\mapsto \underline{g}_{\mathcal{W}}=10,11,12\\
\vdots & \vdots\\
\left\{\left(S_{1,2,3};\begin{bmatrix}0&0\\2&1\\1&2\end{bmatrix}\right);\left(S_{1,2,3};\begin{bmatrix}0&0\\1&2\\2&1\end{bmatrix}\right)\right\}\mapsto & \left(S_{1,2,3};\begin{bmatrix}0\\2\\2\end{bmatrix}\right)\mapsto \underline{g}_{\mathcal{W}}=40,41,42
\end{array}
$$

Table 2. Illustration of the partitioning for the Highest Rank solution

$$
\begin{array}{ll}
\left\{\left(S_{1,2,3};\begin{bmatrix}2&2\\1&1\\0&0\end{bmatrix}\right)\right\}\mapsto & \left(S_{1,2,3};\begin{bmatrix}4\\2\\0\end{bmatrix}\right)\mapsto \underline{g}_{\mathcal{W}}=1,2,3\\
\left\{\left(S_{1,2,3};\begin{bmatrix}2&2\\1&0\\0&1\end{bmatrix}\right);\left(S_{1,2,3};\begin{bmatrix}2&2\\0&1\\1&0\end{bmatrix}\right)\right\}\mapsto & \left(S_{1,2,3};\begin{bmatrix}4\\1\\1\end{bmatrix}\right)\mapsto \underline{g}_{\mathcal{W}}=4,5,6\\
\vdots & \vdots\\
\left\{\left(S_{1,2,3};\begin{bmatrix}2&0\\1&1\\0&2\end{bmatrix}\right);\left(S_{1,2,3};\begin{bmatrix}2&0\\0&2\\1&1\end{bmatrix}\right);\left(S_{1,2,3};\begin{bmatrix}1&1\\2&0\\0&2\end{bmatrix}\right);\right. & \\
\left.\left(S_{1,2,3};\begin{bmatrix}1&1\\0&2\\2&0\end{bmatrix}\right);\left(S_{1,2,3};\begin{bmatrix}0&2\\2&0\\1&1\end{bmatrix}\right);\left(S_{1,2,3};\begin{bmatrix}0&2\\1&1\\2&0\end{bmatrix}\right)\right\}\mapsto & \left(S_{1,2,3};\begin{bmatrix}2\\2\\2\end{bmatrix}\right)\mapsto \underline{g}_{\mathcal{W}}=16,17,18\\
\vdots & \vdots\\
\left\{\left(S_{1,2,3};\begin{bmatrix}0&0\\1&1\\2&2\end{bmatrix}\right)\right\}\mapsto & \left(S_{1,2,3};\begin{bmatrix}0\\2\\4\end{bmatrix}\right)\mapsto \underline{g}_{\mathcal{W}}=55,56,57
\end{array}
$$

Table 3. Illustration of the partitioning for the Borda Count solution

decision is made by choosing the candidate class having the maximum sum score. We consider a partitioning W which is illustrated in Table 3 for an example case of $P = 3$ classes and $Q = 2$ classifiers. Again the usual notation is used. In this partitioning, there are a total of $M_{\mathcal{W}} = 57$ partitions as compared with the original cardinality $P \times N = 108$. The arguments of Fact 1 and 2 and Theorem 1 also hold for the Borda Count method establishing its links with the introduced theory. As discussed in [25], it is also possible to show that the Logistic Regression method introduced first by Ho can be explained as a special case of the POS theory.

7. Conclusions

We have considered the closed-set pattern classification problem and attempted to formulate the rank-based multiple classifier decision problem as a general binary integer programming problem. The optimization has been based on maximizing the total probability of correct decision and has led to an objective function expansion including decision related terms as well as statistics which can be estimated from joint classifier behavior on the cross-validation

data. Although the existence of a global optimum solution is shown, the dimensionality of the problem necessitated techniques of reducing the number of statistics to be estimated. We have proposed the partitioning approach as a controlled tool to selectively achieve dimensionality reduction and provided an illustrative example. The full implications of different partitionings are yet to be explored. However, we have shown that a number of such partitionings establish the relations of the presented theory with some popular rank-based multiple classifier decision methods. We believe that the theory presented is a promising direction for understanding the rank-based multiple classifier decision systems as well as to establish the links between Type 1 and Type 3 systems.

References

[1] B. Achermann and H. Bunke. Combination of face classifiers for person identification. In *Proceedings of IAPR International Conference on Pattern Recognition*, pages 416–420, Vienna, Austria, 1996.

[2] K. Al-Ghoneim and B. Kumar. Learning ranks with neural networks. In *Proceedings of SPIE*, pages 446–464, 1995.

[3] R. Battiti and A. M. Colla. Democracy in neural nets: Voting schemes for classification. *Neural Networks*, 7(4):691–707, 1994.

[4] J. A. Benediktsson, J. R. Sveinsson, and P. H. S. Okan K. Ersoy. Parallel consensual neural networks. *IEEE Transactions on Neural Networks*, 8(1):54–64, January 1997.

[5] J. A. Benediktsson and P. H. Swain. Consensus theoretic classification methods. *IEEE Transactions on Systems, Man and Cybernetics*, 22(4):688–704, July/August 1992.

[6] Y. Bennani and P. Gallinari. Neural networks for discrimination and modelization of speakers. *Speech Communication*, 17:159–175, 1995.

[7] D. Black. *The Theory of Commitees and Elections*. Cambridge University Press, London, 2nd edition, 1963.

[8] R. Brunelli and D. Falavigna. Person identification using multiple cues. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(10):955–966, October 1995.

[9] B. V. Dasarathy. *Decision Fusion*. IEEE Computer Society Press, Los Alamitos, 1994.

[10] M. Demirekler and A. Saranlı. A study on improving decisions in closed set speaker identification. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1127–1130, Munich, Germany, April 1997.

[11] K. R. Farell and R. J. Mammone. Data fusion techniques in speaker recognition. In R. Ramachandran and R. J. Mammone, editors, *Modern Methods of Speech Processing*, chapter 12, pages 279–297. Kluwer Academic Publishers, Boston, Massachusetts, 1995.

[12] K. Fukunaga. *Introduction to Statistical Pattern Recognition*. Academic Press, Boston, 2 edition, 1990.

[13] L. Hansen and P. Salamon. Neural network ensembles. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12:993–1001, 1990.

[14] T. K. Ho. *A Theory of Multiple Classifier Systems and Its Application to Visual Word Recognition*. PhD thesis, Department of Computer Science, State University of New York at Buffalo, May 1992.

[15] T. K. Ho, J. J. Hull, and S. N. Srihari. Decision combination in multiple classifier systems. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(1):66–75, January 1994.

[16] Y. Huang and C. Suen. A method of combining multiple experts for the recognition of unconstrained handwritten numerals. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(1):90–94, January 1995.

[17] F. Kimura and M. Shridhar. Handwritten numerical recognition based on multiple algorithms. *Pattern Recognition*, 24(10):969–983, 1991.

[18] J. Kittler, M. Halef, and R. Duin. Combining classifiers. In *Proceedings of IAPR International Conference on Pattern Recognition*, pages 897–901, Vienna, Austria, August 1996.

[19] J. S.-J. Lee, J.-N. Hwang, D. T. Davis, and A. C. Nelson. Integration of neural networks and decision tree classifiers for automated cytology screening. In *Proceedings of the International Joint Conference on Neural Networks*, volume 1, pages 257–262, Seattle, July 1991.

[20] G. Mani. Lowering variance of decisions using artificial neural networks portfolios. *Neural Computation*, 3:484–486, 1991.

[21] M. P. Perrone and L. N. Cooper. When networks disagree: Ensemble methods for hybrid neural networks. In R. J. Mammone, editor, *Artificial Neural Networks for Speech and Vision*, pages 127–142. Chapman & Hall, London, UK, 1993.

[22] V. Radová and J. Psutka. An approach to speaker identification using multiple classifiers. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1135–1138, Munich, Germany, April 1997.

[23] G. Rogova. Combining the results of several neural network classifiers. *Neural Networks*, 7(5):777–781, 1994.

[24] A. Saranlı. Rank-based multiple-classifier decision combination for speaker identification: Progress on theory and experimental results. Technical Report TR-98-02, Dept. of Electrical and Electronics Engineering, Middle East Technical University, December 1998.

[25] A. Saranlı and M. Demirekler. A unified framework for rank-based multiple classifier decision systems. Submitted to *Pattern Recognition*, March 1999.

[26] D. M. Tax, R. P. Duin, and M. van Breukelen. Comparison between product and mean classifier combination rules. In P. Pudil, J. Novovicova, and J. Grim, editors, *Proceedings of 1st International Conference on Statistical Techniques in Pattern Recognition*, pages 165–170. Institute of Information Theory and Automation, June 1997.

[27] K. Tumer and J. Ghosh. Theoretical foundations of linear and order statistics combiners for neural pattern classifiers. Technical Report TR-95-02-98, Department of Electrical and Computer Engineering, University of Texas, Austin, USA, 1995.

[28] D. H. Wolpert. Stacked generalization. *Neural Networks*, 5:241–259, 1992.

[29] L. Xu, A. Krzyżak, and C. Y. Suen. Methods of combining multiple classifiers and their applications to handwriting recognition. *IEEE Transactions on Systems, Man and Cybernetics*, 22(3):418–435, May/June 1992.