A Unified View of Rank-Based Decision Combination

Abstract

This study presents a theoretical investigation of the rank-based multiple classifier decision problem for closed-set pattern classification. The case with classifier raw outputs in the form of candidate class rankings is considered and formulated as a discrete optimization problem with the objective function being the total probability of correct decision. The problem has a global optimum solution but is of prohibitive dimensionality. 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. The formalism can effectively explain a number of rank-based combination approaches successfully used in the literature one of which is discussed.

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 in a variety of fields including machine printed word/character recognition [4], handwritten character recognition [10, 6], speaker recognition, [3], text to phoneme translation [9], remote sensing [2] and biomedical signal processing [5]. The neural networks community has also been active on this subject [8, 9]. 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 from each classifier, resulting in three categories [10]: The outputs may be single class labels (Type 1), rankings of a subset of classes from highest to lowest "likelihood" (Type 2) and measurement values for the 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 combination of multiple classifiers, using only this abstract level may lead to a loss of valuable information. It should be advantageous to use classifier output forms with more information. It has been shown [4] 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. 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 [4], 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 [1], 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 in [4] with a number of new methods. The *Highest Rank*, *Borda Count* and *Logistic Regression* methods are of special interest for closed-set pattern classification applications. The first two are simple methods which do not use any classifier behavior observation while the last one attempts to generalize the Borda Count method by linearly weighting the classifiers, incorporating the classifier behavior observation into the combination process. The present study also considers the rank-based combination and attempts a theoretical formulation based on discrete optimization and observation space partitioning. The optimality conditions on the three rank-based combination methods can be demonstrated within this unifying formulation.

2. Theoretical Formulation

Consider a closed-set pattern classification problem where patterns belong to P *candidate classes* $S_j, j = 1, 2, \cdots, P$. There are Q classifiers $X_q, q = 1, 2, \cdots, Q$ involved in the classification process. Furthermore, x denotes a *pattern*, causing all classifiers to generate candidate class rankings which are transformed into a *rank score matrix* form $\mathbf{R}$. The elements r_{ji} are positive integer *rank scores* with the highest score assigned to the highest ranking class. We define two random variables taking index values of an ordered set $\mathcal{S}$ of candidate classes: $\underline{s}_x$ denotes the true source class, $\underline{d}$ denotes the final decision of the system. The processing of x by all classifiers results in a rank score matrix $\mathbf{R}$ which is the only input for final classification. Let the objective be to obtain the maximum rate

of correct classification. Other objectives are also possible but this is a meaningful one for closed-set pattern recognition. The total probability of correct classification can be expressed as $P\{\underline{y}=1\}$ where the $\underline{y}$ is a binary valued *indicator* of the correct decision. The problem of finding the best rank-based decision combination process becomes one of maximizing $P\{\underline{y}=1\}$. To be useful, this objective function 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 Bayes rule we obtain

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

By definition, the decision process to be found uses only the rank score matrix, i.e., is a deterministic function of $\underline{r}$. Hence we have $P\{\underline{d}=j|\underline{s}_x=j,\underline{r}=n\} = P\{\underline{d}=j|\underline{r}=n\}$ leading to

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

In this expansion, the first terms $P\{\underline{d}=j|\underline{r}=n\}$ are directly linked with the decision process we are seeking. For a given deterministic decision process, these have uniquely determined binary values "0" and "1". The joint probability terms $P\{\underline{s}_x=j,\underline{r}=n\}$ on the other hand are independent of the decision process and models the joint behavior of the classifier ensemble. This set of probabilities can be estimated if the classifiers are operated on labeled cross-validation data. Denoting the decision terms as our optimization variables b_{jn} and assuming that the *classifier observation statistics* have been properly estimated, we obtain a constrained optimization problem with constraints arising from the fact that there should be a unique decision for a given rank score matrix. That is we have,

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

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

Since all $P\{\underline{s}_x=j,\underline{r}=n\}$ are non-negative, this problem has an obvious global optimum solution given by

$$b^*_{jn} = \begin{cases} 1 & \text{if } j = \underset{k=1,2,\cdots,P}{\operatorname{argmax}} P\{\underline{s}_x=k,\underline{r}=n\}, \\ 0 & \text{otherwise,} \end{cases} \quad (5)$$

2.1. Curse of Dimensionality

The optimal b^*_{jn} correspond to an *optimum decision process*. When an unknown pattern x is processed by all the classifiers, the rank score matrix $\underline{r}$ is determined. The index k of the single non-zero b_{kn} among the P variables corresponding to this $\underline{r}$ is the final classification $\underline{d}=k$. The achieved global optimum decision may not be unique. This occurs when there are more than one maximum coefficient $P\{\underline{s}_x=j,\underline{r}=n\}$ For these cases, the overall system is unable to discriminate among the corresponding classes for that specific rank-score matrix occurrence.

The given solution is possible if we have the observation statistics estimated properly. Unfortunately, there are $P(P!)^Q$ of them, which is prohibitively large for most problems. Since they should be extracted from limited data, a formalism of reducing this dimensionality is required. This can be accomplished by the following formulation.

2.2. 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} (*Problem Parameter Space*), while the second one being the space spanned by the estimated behavior statistics $P\{\underline{s}_x=j,\underline{r}=n\}$(*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. 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 $\mathcal{F}$ which specify 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 $\mathcal{W}$ will denote both the partitioning and the mapping associated with it. Assume that $\mathcal{W}$ results in $M_{\mathcal{W}}$ partitions $W_1, W_2, \cdots, W_{M_{\mathcal{W}}}$ which are disjoint and their union being $\mathcal{F}$. 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 $\underline{g}_{\mathcal{W}} : \mathcal{S}\times\mathcal{R} \longmapsto \{1,2,\cdots,M_{\mathcal{W}}\}$, whose values are indexes on an ordered set $G_{\mathcal{W}} = \{W_1, W_2, \cdots, W_{M_{\mathcal{W}}}\}$. Here, $\mathcal{S}$ is the set of possible source classes while $\mathcal{R}$ is the set of possible rank score matrixes. By observing that the random variable $\underline{g}_{\mathcal{W}}$ is a deterministic mapping from the values of $\underline{s}_x$ and $\underline{r}$, the double sum in (2) can also be written by introducing the new random variable as

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

which, by using the Bayes rule and the fact that the decision should be based on the rank score matrix only, becomes

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

The first and last set of terms inside this expansion have the usual meanings of *decision variables* and *observation statistics*. 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 inside observable partitions. In the middle, we have a set of newly introduced *transition terms* between this coarser resolution and the finer resolution of the original event atoms. Clearly, the first terms will be 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 partition level, there is by definition no data to determine the transition terms. By our partition selection, we are *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*. I.e., we have $P\{\underline{s}_x=j, \underline{r}=n|\underline{g}_{\mathcal{W}}=m\} = 1/|W_m|$, if $(\underline{s}_x=j, \underline{r}=n) \in W_m$ and 0 otherwise, where $|W_m|$ is the cardinality of the partition

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. (For the above expansion we have $M_{\mathcal{W}}$ statistics to estimate.) For fixed cross-validation data, a reduction in the number of statistics to estimate corresponds to an increase in the reliability, which is crucial to the generalization performance, hence to the classification performance of the system [9].

Although we have mentioned that the number of statistics can be reduced, this should be done by considering the amount of data available. The new optimal solution based on 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 important for the usefulness of the resulting solution. The objective should be to maintain the maximum observation resolution which is reasonable for the 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 this will increase the deviation from the global optimum.

Although (7) still has the original number of terms, with the number of estimates reduced to $M_{\mathcal{W}}$, the optimum solution in (5) may be converted into an algorithmic form requiring a small number of computations for making the optimum decision based on the estimated statistics: 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 (7).) I.e., one coefficient is computed for each candidate class. Decide on the class with the maximum coefficient, requiring a total of at most P multiplications. 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.

3. Specific Partitionings

3.1. Example: First Two Ranks

The partitioning rule used to decrease the number of statistics is often task dependent. 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. This will be illustrated with an example *first two ranks based partitioning*. Assume it is intuitively expected that *the rankings below the topmost two ranks (largest two rank scores) are unreliable.* This is a reasonable expectation, e.g., for distance classifiers, since the separation between class models becomes less significant as the models become farther from the unknown input pattern. Hence noisy features have greater affect on the lower rankings. Based on this, we decide not to discriminate among the ranks lower than the second and group them as the *last rank*. This corresponds to a new score assignment and partitioning rule: $\hat{r}_{ij} = r_{ij} - P + 3$, if $r_{ij} > P - 3$ and $\hat{r}_{ij} = 0$ otherwise. For an illustrative example of $P = 4$ classes and $Q = 2$ classifiers, the partitions 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. By the use of a summary notation, each row represents 4 actual partitions corresponding to classes S_1, S_2, S_3, S_4. The contents of each partition consisting of 4 event-atoms are illustrated by a *don't care* notation where the don't care block can take any allowable combination. Actual class names and rank score matrixes are used instead of the random variables for clarity. There are a total of 576 partitions as compared to the original 2304 event atoms.

3.2. Special Case: Highest Rank Method

Now the theory will be linked with one existing rank-based decision combination method by discussing the corresponding partitioning and conditions on optimality. The Highest Rank method [4] is a simple technique which does

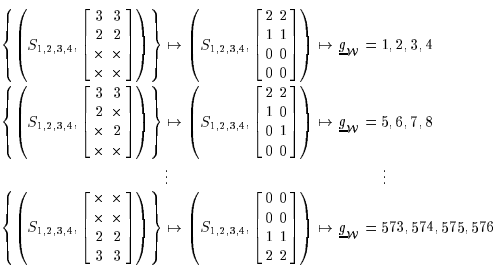

Table 1. First Two Ranks partitioning.

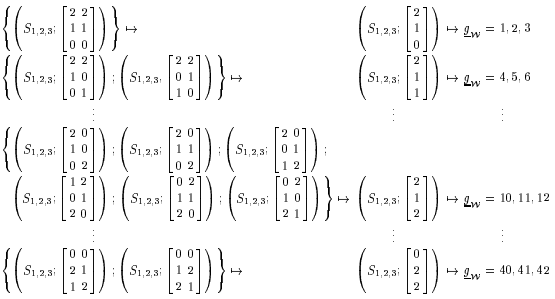

Table 2. Partitioning for Highest Rank Method

not use any estimated behavior model. It is not optimal in the general case. The method can be described as "For each source class, select the highest of the rank scores assigned by all classifiers for that class to form a *max-score* vector. The class with the maximum *max-score* is the decision." *Fact 1*: As a predefined decision process, this method coincides with a specific fixed set of b_{jn} values except when there is max-score collisions, for which we have more than one such set. *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 in (5) using partitioning W and the Highest Rank solution is stated as a theorem whose proof is given elsewhere [7].

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

Table 2 illustrates an example for $P = 3$ and $Q = 2$ the inherent partitioning of the observation space for the Highest Rank Method, by letting each max score vector to correspond to a partition. There are $M_{\mathcal{W}} = 42$ partitions.

Similar line of discussion is possible also for the *Borda Count* and the *Logistic Regression* methods which can be demonstrated to correspond to specific partitionings [7].

4. Conclusion

We have considered the closed-set pattern classification and attempted to formulate the rank-based classifier combination as a discrete optimization problem. The objective function selected as the total probability of correct classification led to an expansion including decision related terms and statistics to be estimated from joint classifier behavior on the cross-validation data. The dimensionality of the problem necessitated techniques of reducing the number of statistics despite the simple global optimum solution. We have proposed a partitioning approach as a controlled tool to selectively achieve dimensionality reduction and provided an example. The full implications of different partitionings are yet to be explored. However, we have argued that a number of partitionings establish the relations of the formalism with some popular rank-based combination methods. We believe that the theory presented is a promising direction for understanding the rank-based systems.

References

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

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

[3] 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.

[4] 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.

[5] Y. H. Hu, S. Palreddy, and W. J. Tompkins. A patient-adaptable ECG beat classifier using a mixture of experts approach. *IEEE Transactions on Biomedical Engineering*, 44(9):891–900, September 1997.

[6] J. Kittler, M. Hatef, R. P. W. Duin, and J. Matas. On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(3):226–239, March 1998.

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

[8] K. Tumer and J. Ghosh. Analysis of decision boundries in linearly combined neural classifiers. *Pattern Recognition*, 29(2):341–348, February 1996.

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

[10] 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.