On Output Independence and Complementariness in Rank-Based Multiple Classifier Decision Systems

Afşar SARANLI* and Mübeccel DEMİREKLER
Speech Processing Laboratory
Department of Electrical and Electronics Engineering
Middle East Technical University, Ankara, Türkiye.
e-mail: {afsars,demirek}@metu.edu.tr

Submitted: February 2000, Revised: October 2000

Abstract

This study presents a theoretical analysis of output independence and complementariness between classifiers in a rank-based multiple classifier decision system in the context of the Partitioned Observation Space theory. To enable such an analysis, an Information Theoretic interpretation of a rank-based multiple classifier system is developed and basic concepts from Information Theory are applied to develop measures for output independence and complementariness. It is shown that output independence of classifiers is not a requirement for achieving complementariness between these classifiers. Namely, output independence does not imply a performance improvement by combining multiple classifiers. A condition called Dominance is shown to be important instead. The information theoretic measures proposed for output independence and complementariness are justified by simulated examples.

Keywords: statistical classifier combination, statistical decision combination, statistical pattern recognition, multiple classifier systems, ranks, classifier observation space, event space partitioning, Bayesian formalism, independence, complementariness, entropy, mutual information.

1 Introduction

Multiple classifier systems have been a focus of intensive research for the last decade. Contributions have been made or some form of decision combination system have been attempted in a variety of pattern recognition fields. These include machine printed word/character recognition [1], handwritten character recognition [2, 3, 4, 5, 6, 7, 8], speaker recognition, [9, 10, 11, 12], face identification [13, 14], text to phoneme translation [15], remote sensing [16, 17], military target

*Corresponding author. E-mail: afsars@metu.edu.tr

recognition [18] and biomedical signal processing [19, 20]. The neural networks community has also been active on this approach [21, 5, 22, 15, 23, 24, 25, 6]. Xu and his colleagues have categorized multiple classifier decision combination systems with respect to the type of raw output information from each classifier [2], resulting in three categories: The classifier outputs may be single class labels (Type 1), rankings of a subset of source classes from highest to lowest "likelihood" (Type 2 or *rank-based*) or the complete set of *similarity score* values for the candidate classes leading to such rankings (Type 3).

Two closely related concepts arise while using a multiple-classifier system with the aim of improving the overall classification performance. These are the *independence* and the *complementariness* of the classifiers involved.

While constructing a multiple classifier decision combination system, one is faced with several important problems. It is often not clear whether there will really be an improvement over the performance of the best classifier by the use of more than one classifier. This clearly depends on the individual performances of the classifiers involved and their interaction during the classification process. One is faced with the problem of determining the potential improvement possible by making collective use of multiple classifiers.

Another issue of considerable importance is the computational load implied by the parallel use of multiple classifiers. Given a large set of potential classifiers (e.g., using different features extraction methods, different modeling/similarity scoring methods), using all of them in parallel may guarantee performance improvement but may not be computationally feasible with the hardware capabilities at hand. Practical considerations often necessitate selecting a suitable subset of classifiers which satisfy a certain performance gain/classification speed tradeoff.

Finally, one should be interested in understanding theoretically *when* and *why* a given set of classifiers, when combined, lead to improved performance, while others do not. Loosely stated, the aforementioned objectives may only be achieved if it is possible to quantify the potential of the combiner to improve the classification performance. The *complementariness* concept and an associated measure may be used to quantify such an ability.

Independence and *complementariness* concepts have been around in the pattern recognition literature for a long time. Unfortunately, the concepts have been often used loosely, without

any attempt for a solid definition and the development of a quantifying measure. For example, the dependence between the set of classifiers is often ignored and a statistical independence assumption is used in the development [22, 7]. Some other researchers have argued that statistical independence of the classifier outputs is not really the useful measure for quantifying improved performance but the *independence of the errors made* should be considered instead. This is also left as a verbal argument [5, 6]. There have also been solid contributions such as by Tumer and Ghosh [21, 21, 26]. They have shown the relations between classifier output correlation and the deviation from the optimal Bayesian decision boundary for classifiers which are combined by linear averaging or by order statistics. Their results apply to classifiers with continuous outputs in measurement form and cannot be extended trivially to rank-based classifier systems.

Recently, the authors have introduced the *Partitioned Observation Space* (POS) Theory as a unifying view where all rank-based multiple classifier systems are uniformly treated and the decision combination problem is formulated as one of discrete optimization [27, 28]. The rank-based multiple classifier system is treated as an interrelated set of random variables and the partitioning of the classifier observation space is introduced as a controlled tool to selectively reduce the observation resolution in order both to reduce the problem dimensionality and to suppress undesired or unreliable resolution.

In the present study, building on the concepts developed in [27] and some basic concepts from Information Theory, a formal treatment of classifier *independence* and *complementariness* concepts for rank-based Type 2 multiple classifier systems will be attempted. For this purpose, first a definition will be proposed for the output independence of rank-based classifiers. It will be argued that this Information Theory based definition also gives an output dependence measure for the classifiers involved. Then the improvement in performance by combining the rank outputs of more than one classifiers is questioned from the output independence point of view. It is shown by an example that dependent classifiers when combined may give a better performance than the combination of independent classifiers if the combined decision is done in an optimal way in the Bayesian sense. This observation leads to a deeper analysis of the conditions for improvement and finally to the definition of *dominance* between combined classifiers and a proposal for a measure of *complementariness* between them.

The paper is organized along the aforementioned ideas as follows. The Partitioned Observation Space Theory is presented in Section 2. In Section 3, some relevant concepts from Information Theory are introduced. Then in Sections 4 and 5, the paper develops the information theoretic interpretation of a multiple classifier system and discusses the output independence of classifiers and its relations with complementariness. In Sections 6 and 7, first a necessary and sufficient condition on complementariness and then an information theoretic measure is developed. The proposed measures are justified by means of illustrative examples and the paper concludes in Section 8 by a discussion of these theoretical results.

2 The Partitioned Observation Space Theory

Consider a closed-set pattern classification problem where patterns belong to P *source 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 source 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 source 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. Possible rank score matrixes are denoted by yet another index random variable $\underline{r}$ such that $(\underline{r} = n)$ denotes the realization of the rank score matrix $\mathbf{R}_n$ on a finite event space $\mathcal{R} = \{\mathbf{R}_1, \mathbf{R}_2, \cdots, \mathbf{R}_N\}$. 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 $\underline{y}$ is a binary valued *indicator* of the correct decision, which is "1" for correct classification and "0" otherwise. 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

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

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

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

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 joint probabilities 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,

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

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

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

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

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 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 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 of the observation statistics. 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 can be summarized as follows [27].

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

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

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

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

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 [15].

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. A number of sensible partitionings are discussed in [27] and [29] where some specific partitionings are shown to lead to existing methods from the rank-based classifier combination literature.

The optimum solution to the optimization problem given in (7) is similar to the solution to the original problem, but with the number of estimates now reduced to $M_{\mathcal{W}}$. This solution may be applied in an algorithmic form requiring a small number of computations for making the optimum decision based on the estimated statistics: For any given rank score matrix ($\underline{r} = n$), the coefficients $P\{\underline{s}_x = j, \underline{r} = n | \underline{g}_{\mathcal{W}} = \mathcal{W}(j,n)\} \cdot P\{\underline{g}_{\mathcal{W}} = \mathcal{W}(j,n)\}$ must be considered for $j = 1, 2, \cdots, P$. The index j of the largest of these coefficients determines the final class decision of the system. This procedure requires 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 Relevant Concepts of Information Theory

Information theory gives us a promising tool to explore the complementariness of multiple classifiers. To illustrate this, we will first summarize some relevant basic results using the notation of Section 2 [30].

Consider again the finite event space $\mathcal{R} = \{\mathbf{R}_1, \mathbf{R}_2, \cdots, \mathbf{R}_N\}$ and let $\underline{r}$ be an integer valued random variable as defined in Section 2.2. This event space can be thought as a source of

information. One can define a *measure* for the information conveyed by the realization of the event $(\underline{r} = n)$ in terms of its probability as

$$I(\underline{r} = n) = \log \frac{1}{P\{\underline{r} = n\}}. \tag{8}$$

The expected value of the information acquired by the observation of $\mathcal{R}$ is

$$\begin{aligned} H(\underline{r}) &= E\{I(\underline{r} = n)\} \\ &= \textstyle\sum_{n=1}^{N} P\{\underline{r} = n\} \log \frac{1}{P\{\underline{r}=n\}}, \end{aligned} \tag{9}$$

which is also known as the *entropy* of this information source. This quantity can be interpreted as a number of properties of the event space $\mathcal{R}$ or the associated random variable $\underline{r}$ [30]. These are *the amount of average "information" conveyed by an observation of* $\underline{r}$, *our uncertainty about* $\underline{r}$ or *the randomness of* $\underline{r}$. The units of these information measures depend on the base of the log(.) operator. For a base 2 logarithm, the unit of information is *bit*. The well known Theorem 1 establishes the minimum and maximum values for the entropy function and its proof can be found in [30].

THEOREM 1 *Let* $\underline{r} \in \{1, 2, \cdots, N\}$; *then one has* $0 \leq H(\underline{r}) \leq \log N$. *Furthermore* $H(\underline{r}) = 0$ *iff* $\exists j \in \{1, 2, \cdots, N\}$ *such that* $P\{\underline{r} = j\} = 1$ *and* $H(\underline{r} = \log N)$ *iff* $\forall j \in \{1, 2, \cdots, N\}$ *we have* $P\{\underline{r} = j\} = 1/N$.

Consider now that there are two random variables $\underline{r}_1$ and $\underline{r}_2$ with probability mass functions $P\{\underline{r}_1 = n_1\}$ and $P\{\underline{r}_2 = n_2\}$, representing two related event spaces $\mathcal{R}_1$ and $\mathcal{R}_2$. The relation between these two probability distributions is given by the the conditional probability $P\{\underline{r}_1 = n_1 | \underline{r}_2 = n_2\}$. Now if we define the information conveyed by observing the realization $(\underline{r}_1 = n_1)$ given that we have already observed $(\underline{r}_2 = n_2)$ as

$$I(\underline{r}_1 = n_1 | \underline{r}_2 = n_2) = \log \frac{1}{P\{\underline{r}_1 = n_1 | \underline{r}_2 = n_2\}}, \tag{10}$$

then the *entropy* of $\underline{r}_1$ after observing $\underline{r}_2$ can be found [30] as

$$H(\underline{r}_1 | \underline{r}_2) = \sum_{n_1, n_2} P\{\underline{r}_1 = n_1, \underline{r}_2 = n_2\} \log \frac{1}{P\{\underline{r}_1 = n_1 | \underline{r}_2 = n_2\}}. \tag{11}$$

This *conditional entropy* may be interpreted as a number of properties of $\underline{r}_1$ and $\underline{r}_2$: *The amount of average "information" conveyed by an observation of* $\underline{r}_1$ *given that we have already observed* $\underline{r}_2$*, our uncertainty remaining about* $\underline{r}_1$ *given that we have resolved our uncertainty about* $\underline{r}_2$ or *the randomness of* $\underline{r}_1$ *after observing* $\underline{r}_2$. Since we know our uncertainty about $\underline{r}_1$ both *before* and *after* observing $\underline{r}_2$, we can derive the amount of average information we have acquired about the former by observing the latter. This symmetric quantity is known as the *mutual information* between $\underline{r}_1$ and $\underline{r}_2$ and is given by

$$I(\underline{r}_1, \underline{r}_2) = H(\underline{r}_1) - H(\underline{r}_1|\underline{r}_2). \tag{12}$$

which can be expressed in explicit form as

$$I(\underline{r}_1, \underline{r}_2) = \sum_{n_1, n_2} P\{\underline{r}_1 = n_1, \underline{r}_2 = n_2\} \log \frac{P\{\underline{r}_1 = n_1, \underline{r}_2 = n_2\}}{P\{\underline{r}_1 = n_1\} P\{\underline{r}_2 = n_2\}}. \tag{13}$$

THEOREM 2 *We have* $I(\underline{r}_1, \underline{r}_2) \geq 0, \forall \underline{r}_1, \underline{r}_2$ *and* $I(\underline{r}_1, \underline{r}_2) = 0$ *if and only if the two random variables are statistically independent.*

Theorem 2 whose proof can be found in [30] asserts that the mutual information as defined in(12) is a well suited measure of statistical dependence between the random variables $\underline{r}_1$ and $\underline{r}_2$ hence between the underlying events [30]. These concepts can be applied in the context of multiple classifier systems as discussed in the following sections.

4 An Information Theoretic Interpretation of Classifiers

Information theory defines a *discrete memoryless communication channel* (DMC) as *an object that accepts, every unit of time, one of* P *input symbols and outputs one of* N *output symbols.* The output can be thought of as a noisy version of the input [30]. A classifier on the other hand, is an object which accepts patterns, whose class labels are known to a *supervisor*, and outputs its best estimates of these class labels.

A classifier can be interpreted as analogous to a DMC if we argue that the true realization of the class label is transformed by the classifier into a noisy output form. The source of the noise is not important for this interpretation but it may be the result of the feature extraction

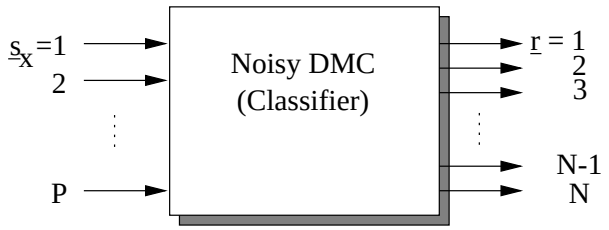

Figure 1: The discrete memoryless channel interpretation of a classifier. The input to the DMC is the true label of the pattern while the output of the DMC is the classifier output. The exact number of outputs depends on the level of information supplied by the classifier.

and/or the similarity scoring algorithm. The actual source of information we are interested in (the input to the DMC interpretation of the classifier) is the true label of the class emitting the patterns. However, what we have access to is only the noisy output of this DMC as illustrated by Figure 1.

When more than one classifiers are involved, we may consider them as multiple DMCs transmitting the same information source whose outputs are to be considered to acquire information about this source.

5 Output Independence of Classifiers

A multiple classifier decision combination system with observation space partitioning can be visualized as a set of interrelated random variables as illustrated in Figure 2. With the random variable definitions given in Figure 2, we are at a point to introduce a formal definition of independence among the outputs of classifiers both before and after observation space partitioning as described in Section 2. Consider two classifiers whose rank-based outputs represented by the random variables $\underline{r}_1, \underline{r}_2$. In view of Theorem 2 we can make the following definition which can easily be extended to more than two classifiers.

DEFINITION 1 *Classifiers X_1, X_2 are said to be* output independent in the rank-based sense *if and only if we have $I(\underline{r}_1, \underline{r}_2) = 0$ with $I(\underline{r}_1, \underline{r}_2)$ defined by*

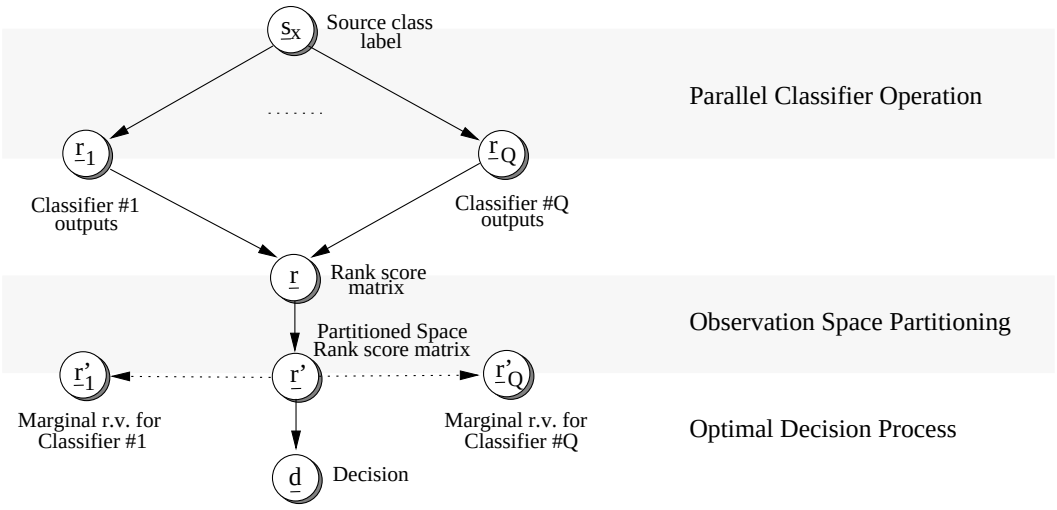

Figure 2: Random variable representation of the multiple classifier decision combination system. The events within the system can be represented by a number of interrelated random variables. The random variables are transformed from one to another either by means of the classifiers, or by means of the partitioning and the optimal decision process. $\underline{r}'_1, \underline{r}'_2, \cdots, \underline{r}'_Q$ are the marginal random variables reflecting the individual classifier outputs after the application of the observation space partitioning.

$$I(\underline{r}_1, \underline{r}_2) = \sum_{n_1,n_2,j} P\{\underline{r}_1 = n_1, \underline{r}_2 = n_2 | \underline{s}_x = j\} \log \frac{P\{\underline{r}_1=n_1,\underline{r}_2=n_2|\underline{s}_x=j\}}{P\{\underline{r}_1=n_1|\underline{s}_x=j\}P\{\underline{r}_2=n_2|\underline{s}_x=j\}}. \quad (14)$$

Otherwise, the two classifiers are output dependent *with* $I(\underline{r}_1, \underline{r}_2)$ *being a measure of dependence between them.*

If one uses the random variables $\underline{r}_1, \underline{r}_2, \cdots, \underline{r}_Q$ in this definition, then the output dependence of the original classifiers is computed. However, it is also possible to compute the output dependence, after a partitioning of the observation space as described in Section 2. For this, the marginal output random variables $\underline{r}'_1, \underline{r}'_2, \cdots, \underline{r}'_Q$, derived after the partitioning $\mathcal{W}$, should be used instead of $\underline{r}_1, \underline{r}_2, \cdots, \underline{r}_Q$. Note that the numerical measure of dependence among the outputs of the classifiers will be different depending on whether this is computed for the original outputs or after each specific partitioning

X_1X_2	$\begin{bmatrix}1&1\\0&0\end{bmatrix}$	$\begin{bmatrix}1&0\\0&1\end{bmatrix}$	$\begin{bmatrix}0&1\\1&0\end{bmatrix}$	$\begin{bmatrix}0&0\\1&1\end{bmatrix}$
S_1	0.45	0.35	0.15	0.05
S_2	0.15	0.05	0.05	0.75

Table 1: True joint probability distribution of the classifier observation space. Columns denote the *rank score matrixes* while rows denote pattern classes. Each cell represent the estimate of the probability that patterns from a class lead to a specific rank score matrix at the outputs of the classifiers.

Output independence of classifiers is an important parameter in itself. However, as is shown by the following example, it is not necessarily a measure of *complementariness*.

EXAMPLE 1 Suppose we consider two rank-based classifiers X_1 and X_2 operating on a simple two class problem where the class labels are S_1 and S_2. Assume that these classifiers are operated in parallel on patterns from these two classes and the class conditional joint probabilities in Table 1 are obtained. [1] These will be called as the *true* joint distribution of the classifier behavior. The marginal probabilities for the individual classifiers can be obtained from this joint distribution and are given in Table 2 (a) and (b).

Probability of the errors made by the individual classifiers may be analyzed from these two marginal tables. Considering Table 2 (a), the jointly optimum decision [2] selects class S_1 if the rank score matrix $(1\ 0)^T$ occurs and S_2 if $(0\ 1)^T$ occurs at the classifier outputs. Denoting the decision by $\underline{d}$ and using the random variable notations of Section 2, the total probability of error for classifier X_1 is

$$\begin{aligned} P^e_{X_1} &= P\{\underline{d}=1|\underline{s}_x=2\}P\{\underline{s}_x=2\} + P\{\underline{d}=2|\underline{s}_x=1\}P\{\underline{s}_x=1\} \\ &= 0.2\times 0.5 + 0.2\times 0.5 \\ &= 0.2 \end{aligned}$$

[1] From now on, we will drop mentioning that the probabilities are conditional except for the cases where there is an ambiguity.

[2] The *jointly optimum decision* is in the sense of Section 2. In this sense, the *jointly optimum decision* and the *optimum combination* is synonymous. For this example, the class label with the largest probability for a given column is selected.

(a)

X_1	$(1\ 0)^T$	$(0\ 1)^T$
S_1	0.8	0.2
S_2	0.2	0.8

(b)

X_2	$(1\ 0)^T$	$(0\ 1)^T$
S_1	0.6	0.4
S_2	0.2	0.8

Table 2: Marginal probabilities for individual classifiers (a) X_1 and (b) X_2. Columns denote the *rank score vectors* at classifier outputs while rows denote pattern classes. These tables are rank-based generalized forms of the matrixes known as classifier *confusion matrixes*.

X_1X_2	$\begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix}$	$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$	$\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$	$\begin{bmatrix} 0 & 0 \\ 1 & 1 \end{bmatrix}$
S_1	0.48	0.32	0.12	0.08
S_2	0.04	0.16	0.16	0.64

Table 3: Joint probability distribution of the classifier observation space computed from the marginal distributions in Table 2, under the *independence assumption*.

By a similar computation for classifier X_2, we have $P^e_{X_2} = 0.3$. Therefore, it can be argued that X_1 is the best of the two classifiers.

Let the true joint distribution in Table 1 be used for jointly optimal decision. From this table, it can be seen that one has $\underline{d} = 1$ if $\underline{r} \in \{1, 2, 3\}$ and $\underline{d} = 2$ if $\underline{r} = 4$, where $\underline{r}$ denotes the realization of the rank score matrix. The total probability of error for the jointly optimal decision is

$$\begin{aligned} P^e_{X_1X_2} &= (0.15 + 0.05 + 0.05) \times 0.5 + 0.05 \times 0.5 \\ &= 0.15 \end{aligned}$$

which is lower than the probability of error $P^e_{X_1} = 0.2$ for the best individual classifier. Therefore, an improvement in performance over the best individual classifier is achieved by the jointly optimal decision. Now suppose that the classifiers are independent. Then, we can construct a joint probability distribution by making use of this assumption. This derived distribution is given in Table 3.

When this derived joint distribution is considered for optimal decision, one has now $\underline{d} = 1$ if $\underline{r} \in \{1, 2\}$ and $\underline{d} = 2$ if $\underline{r} \in \{3, 4\}$. In this case, the total probability of error would clearly be

X_1X_2	$\begin{bmatrix}1&1\\0&0\end{bmatrix}$	$\begin{bmatrix}1&0\\0&1\end{bmatrix}$	$\begin{bmatrix}0&1\\1&0\end{bmatrix}$	$\begin{bmatrix}0&0\\1&1\end{bmatrix}$
S_1	0.12	0.48	0.08	0.32
S_2	0.05	0.45	0.05	0.45

Table 4: Joint probability distribution of the classifier observation space for the two classifiers of Example 2.

$$\begin{aligned}\hat{P}^e_{X_1X_2} &= (0.04+0.16)\times 0.5+(0.12+0.08)\times 0.5\\ &= 0.20,\end{aligned}$$

which shows no improvement over the performance of the best classifier X_1.

This simple example shows that the independence assumption may hide a potential for improvement for classifiers which are in fact dependent. It also shows that independence of classifiers is not a necessary condition for such an improvement. For dependent classifiers, the jointly optimal decision process in the sense of the theory summarized in Section 2 may achieve an improvement over the best individual classifier while methods based on the independence assumption will fail to do so. An interesting question at this point is whether or not an improvement is still possible for the case of classifiers which are truly output independent. The following example gives a positive answer.

EXAMPLE 2 Again consider a simple problem with two classifiers X_1 and X_2, operating on patterns from two classes S_1 and S_2. The joint distribution of the classifier observation space is given in Table 4 while the marginal distributions for the individual classifiers are given in Table 5 (a) and (b).

For this example, we have $I(\underline{r}_1,\underline{r}_2))=0$ and therefore, the classifiers are output independent. The total probability of error for both individual classifiers are $P^e_{X_1}=P^e_{X_2}=0.45$. However, when the joint distribution is considered for optimal decision, the decisions are $\underline{d}=1$ when $\underline{r}\in\{1,2,3\}$ and $\underline{d}=2$ when $\underline{r}=4$ effectively leading to a total probability of error of $P^e_{Comb}=0.435$. This is smaller than the probability of error for both of the classifiers denoting an improved performance for the case of output independent classifiers.

(a)

X_1	$(1\ 0)^T$	$(0\ 1)^T$
S_1	0.6	0.4
S_2	0.5	0.5

(b)

X_2	$(1\ 0)^T$	$(0\ 1)^T$
S_1	0.2	0.8
S_2	0.1	0.9

Table 5: Marginal probabilities for individual classifiers (a) X_1 and (b) X_2 in Example 2.

Classifier	X_1	X_2
$P\{\underline{y} = 0 \mid \underline{s}_x = 1\}$	0.4	0.8
$P\{\underline{y} = 0 \mid \underline{s}_x = 2\}$	0.5	0.1

Table 6: Class dependent error probabilities for classifiers in Example 2.

An interesting observation can be made about these classifiers if one inspects the *class dependent error probabilities* $P\{\underline{y} = 0|\underline{s}_x = 1\}$ and $P\{\underline{y} = 0|\underline{s}_x = 2\}$ where $\underline{y}$ is the *indicator* of correct decision as defined in Section 2. These are given in Table 6. From these probabilities, it can be concluded that classifier X_1 cannot successfully classify patterns from class S_2 while classifier X_2 cannot classify patterns from class S_1. The fact that the errors of the two classifiers are concentrated on different classifiers support the ideas in [5, 6].

6 A Condition for Complementariness

The joint distribution given in Table 3 is obtained from the marginal distributions under the assumption of independence. However, this could as well have been the true joint distribution of the classifier observation space. Given the true joint distribution and the marginal distributions, one important task is to find the conditions on these distributions so that there will be an improvement by using the jointly optimal decision. Such a general condition is introduced by the following Definition and Fact.

DEFINITION 2 *In a multiple classifier system, a classifier is called as the* dominating classifier *if the jointly optimal decision is a function of only the rank score vector of that classifier.*

FACT 1 *If one classifier dominates the others, then the jointly optimal performance of the multiple classifier system becomes exactly equal to the performance of the dominating classifier.*

The truth of Fact 1 is intuitively apparent from the Definition but a proof can be found in [29]. This fact shows for the general case that if one classifier dominates the others, no improvement can be expected from the combination of the classifiers. Conversely, for improvement by combination, no classifier should dominate, i.e., the jointly optimal decision should favor each classifier's decision in turn, for some rank score matrixes. This is expressed by Theorem 3 which makes use of Lemma 1.

LEMMA 1 *Due to the joint optimality of the combined decision, the combined performance cannot be lower than the performance of the best classifier within a multiple classifier system.*

PROOF. To show this, assume, without loss of generality that X_1 is the best individual classifier. First define $\mathcal{R}_1^{X_1}$ as the set of rank score *vectors* for which the single classifier X_1 decides on class label S_1. Also let Ω be the set of all allowable rank score *vectors* for this single classifier. Now define $\mathcal{R}_1^C$ as the set of all rank score *matrixes* for which X_1 decides on class label S_1, as given by

$$\mathcal{R}_1^C = \left\{ \mathbf{n} = [n_1 \; n_2 \; \cdots \; n_Q] \,|\, n_1 \in \mathcal{R}_1^{X_1}, n_k \in \Omega \; ; k = 2, \cdots, P \right\}. \tag{15}$$

Again without loss of generality, the rank score matrixes can be ordered such that these $L = |\mathcal{R}_1^C|$ rank score matrixes correspond to the random variable values $\underline{r} = 1, 2, \cdots, L$. The corresponding part of the joint distribution of the observation space is illustrated in Figure 3. If the conditions $p_{1n} > p_{jn}$ for $j = 2, 3, \cdots, P$ and $n = 1, 2, \cdots, L$ are satisfied, then the jointly optimal decision is equivalent to the decision of X_1 for this set of $\underline{r}$ values.

Suppose we try to disturb this condition by letting $p_{k1} > p_{11}$ for the $\underline{r} = 1$. This largest probability term will contribute to the probability of error made by X_1. However, it will not contribute to the probability of error made by the optimal decision since the optimal decision will select S_k for $\underline{r} = 1$. Therefore, the error for the optimum decision will necessarily be *lower* than the error for the best classifier X_1. □

THEOREM 3 *If none of the classifiers in a Q classifier ensemble dominate the ensemble, then we necessarily have $P_{Comb}^e < \min\{P_{X_1}^e, P_{X_2}^e, \cdots, P_{X_Q}^e\}$.*

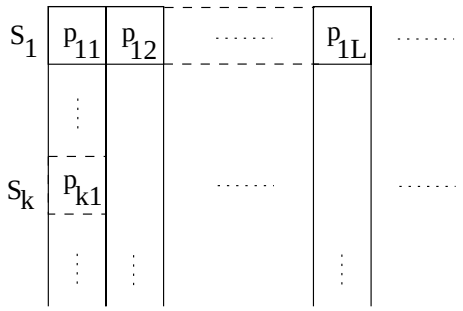

Figure 3: Part of the joint distribution of the observation space used for Lemma 1.

PROOF. Without loss of generality, assume that classifier X_1 is the best performing individual classifier. However, it is not a dominating classifier since there is none. Define $\mathcal{D}^q(\mathbf{r}_q^l)$ to represent the decision of classifier X_q for the specific rank score vector $\mathbf{r}_q^l$ while $\mathcal{D}^C(\mathbf{R}_l)$ denotes the jointly optimal decision for the specific rank score matrix $\mathbf{R}_l = [\mathbf{r}_1^l\ \mathbf{r}_2^l\ \cdots\ \mathbf{r}_Q^l]$.

The fact that X_1 is not a dominating classifier means that there exist at least one or more rank score matrixes $\mathbf{R}_l$ such that $\mathcal{D}^C(\mathbf{R}_l) \neq \mathcal{D}^1(\mathbf{r}_1^l)$. For each such rank score matrix, an intermediate, *partially optimal* decision process $\hat{\mathcal{D}}^l$ can be designed which satisfies $\hat{\mathcal{D}}^l(\mathbf{R}_l) = \mathcal{D}^C(\mathbf{R}_l)$ while for all other rank score matrixes its decisions coincide with the decision of classifier X_1, i.e., $\hat{\mathcal{D}}^l(\mathbf{R}_k) = \mathcal{D}^1(\mathbf{r}_1^k), \forall \mathbf{R}_k \in \mathcal{R},\ k \neq l$. By Lemma 1, the partially optimized decision process cannot yield a performance lower than the performance of the best individual classifier. Therefore, such a decision process which is *different* than the best individual classifier should necessarily yield to an improved performance. □

Another result of this section about dominance is given by Corollary 1.

COROLLARY 1 *If there is a dominating classifier within a multiple classifier system, then this is necessarily the best performing individual classifier.*

PROOF. By Fact 1, the performance of the dominating classifier equals the performance of the combination. However, by Lemma 1, the performance of the combination cannot be lower than the best individual performance. Therefore, the performance of the dominating classifier equals the performance of the best classifier, proving the Corollary. □

The above discussion suggests that output independence plays no exclusive role in assessing the potential for improvement by the combination of classifiers. However, a different concept the paper defines as the *dominance of a classifier* gives a condition on classifier complementariness. Namely, one should have no dominating classifier in a given classifier ensemble in order to have performance improvement by optimal combination in the sense of Section 2.

7 Complementariness of Classifiers

The previous section defined a condition for achieving complementary behavior among classifiers and hence, to obtain an improvement from classifier combination. However, the fact that none of the classifiers are *dominating*, does not give one, a measure on the potential improvement possible by the combination of a set of classifiers. In the present section, an attempt is made to introduce such a measure.

Consider again Figure 2. Apart from the probability of correct classification, another measure on the performance of an individual classifier X_k may be given by means of the mutual information $I(\underline{r}_k, \underline{s}_x)$ between the classifier output $\underline{r}_k$ and the source class $\underline{s}_x$, i.e., it may be argued that the *amount of information acquired about the true class label by observing the outputs of classifier* X_k is a reasonable measure on that classifier's performance.

Now consider that while using X_k individually, one asks the question: *How much does classifier* X_l *has a potential to complement the present classifier* X_k*?*. This depends on the ability of X_l to provide *additional* information about the source class label. Namely, one should be interested in *the amount of new information provided by the output of* X_l *which was not present in the output of* X_k. This quantity can be expressed as a difference

$$\Delta I_{X_k X_l} \doteq I(\underline{r}_k, \underline{r}_l; \underline{s}_x) - I(\underline{r}_k, \underline{s}_x), \tag{16}$$

where the first term represents the amount of information acquired about the source class label $\underline{s}_x$ by observing both classifier outputs $\underline{r}_k$ and $\underline{r}_l$ while the last term represents the amount of information acquired about the source class label by observing the output of classifier X_k alone. Replacing both mutual information terms by their entropy definitions as given in (12) one gets

$$\Delta I_{X_k X_l} = H(\underline{s}_x|\underline{r}_k) - H(\underline{s}_x|\underline{r}_k, \underline{r}_l). \tag{17}$$

which can be expressed in expanded form as

$$\Delta I_{X_k X_l} = \sum_{j,n_1,n_2} P\{\underline{s}_x = j, \underline{r}_1 = n_1, \underline{r}_2 = n_2\} \log \frac{P\{\underline{s}_x = j|\underline{r}_1 = n_1, \underline{r}_2 = n_2\}}{P\{\underline{s}_x = j|\underline{r}_1 = n_1\}}. \tag{18}$$

The quantity we have defined in (16) is not symmetric, namely, we have $\Delta I_{X_k X_l} \neq \Delta I_{X_l X_k}$. This is a reasonable behavior since for classifiers with different performances, the amount of information contributed by X_l to X_k cannot be the same as the amount contributed by X_k to X_l. One expects the contribution of the better performing classifier to be larger.

The quantity defined by $\Delta I_{X_k X_l}$ can be proposed as a measure of the complementariness of classifier X_l with respect to classifier X_k. This proposal is supported by investigating the behavior of the aforementioned measures on several examples with two classifiers and two classes. The joint distributions for these five examples are chosen so as to illustrate some interesting cases of behavior with respect to increasing complementariness, and the associated behavior of the proposed measures. The cases are also selected in such a way that the marginal classifier observation space distributions for classifier X_1 and hence the associated performances are always the same while they are different for the second classifier X_2. Other than these considerations, the joint distributions are arbitrary and do not reflect the behavior of any particular type of classifiers. The distributions and their derived marginal distributions are given in Table 7. Three of these distributions can be recognized from Examples 1 and 2. The given cases are also selected in such a way that the performance and the marginal classifier observation space distribution for classifier X_1 is always the same, while they vary for the second classifier X_2.

Consider the following scenario while investigating Tables 7 and 8. One is restricted to use only two classifiers in parallel for this two class illustrative problem. Five different classifiers are available and the best classifier is labeled X_1. The task is to select the second classifier X_2 among the available ones which is the *most complementary* with respect to the best classifier X_1, i.e., the largest performance improvement over the performance of the best classifier is sought. For this purpose, each alternative classifier is operated in parallel with the best one and the distributions in Table 7 are obtained. From these distributions, the measures in Table 8 are obtained where

	Joint				X_1		X_2	
Case 1	0.48	0.32	0.12	0.08	0.80	0.20	0.60	0.40
	0.04	0.16	0.16	0.64	0.20	0.80	0.20	0.80
Case 2	0.51	0.29	0.09	0.11	0.80	0.20	0.60	0.40
	0.12	0.08	0.08	0.72	0.20	0.80	0.20	0.80
Case 3	0.70	0.10	0.07	0.13	0.80	0.20	0.77	0.23
	0.08	0.12	0.11	0.69	0.20	0.80	0.19	0.81
Case 4	0.66	0.14	0.14	0.06	0.80	0.20	0.80	0.20
	0.09	0.11	0.11	0.69	0.20	0.80	0.20	0.80
Case 5	0.45	0.35	0.15	0.05	0.80	0.20	0.60	0.40
	0.15	0.05	0.05	0.75	0.20	0.80	0.20	0.80

Table 7: Five simulated example cases. Joint and individual classifier observation space distributions are illustrated as three columns. Measures computed from these distributions are given in Table 8.

	Ind. Perf.		Indiv. Infor.		Independence	Joint Infor.	Complementariness		Improv.
Case	$P^e_{X_1}$	$P^e_{X_2}$	$I(\underline{r}_1, \underline{s}_r)$	$I(\underline{r}_2, \underline{s}_r)$	$I(\underline{r}_1, \underline{r}_2)$	$I(\underline{r}_1, \underline{r}_2; \underline{s}_r)$	$\Delta I_{X_1X_2}$	$\Delta I_{X_2X_1}$	ΔP^e_{Comb}
Case 1	0.2	0.3	0.2781	0.1245	0.0000	0.3888	0.0807	0.2343	0.000
Case 2	0.2	0.3	0.2781	0.1245	0.0846	0.3204	0.0423	0.1959	0.005
Case 2	0.2	0.21	0.2781	0.2591	0.1008	0.3592	0.0811	0.1001	0.010
Case 4	0.2	0.2	0.2781	0.2781	0.0359	0.4033	0.1252	0.1252	0.015
Case 5	0.2	0.3	0.2781	0.1245	0.1538	0.4319	0.1538	0.3073	0.050

Table 8: Intermediate measures of interest for the examples with two classes and two classifiers, given in Table 7. The complementariness of classifier X_2 with respect to classifier X_1 is given in the column labeled as $\Delta I_{X_1X_2}$ and is the primary measure of interest.

all logarithms are Base 2 logarithms. This gives a measurement unit of *Bits*. One can make the following discussions.

For this two class problem with uniform class distribution, the entropy of the source random variable $\underline{s}_x$ is 1 bit, which is hence the maximum value for all measures in Table 8 based on Information Theory. For Case 1, the best classifier is dominating the pair since the optimal decision on the joint distribution is the same as the decision of the best classifier X_1 for all cases. Therefore, the candidate classifier cannot contribute to the best classifier and so there is no performance improvement. However, it is interesting to note that the $\Delta I_{X_1X_2}$ column still reports a positive value. It can be argued that the dominance condition may not be reflected in $\Delta I_{X_1X_2}$.

For the remaining cases which are ordered with respect to the actual performance improvement over the best, the best classifier is not dominating. Also, the $\Delta I_{X_1X_2}$ column seems to reflect the potential improvement achievable by combination. Investigating the output independence column $I(\underline{r}_1, \underline{r}_2)$ supports that output independence is not necessarily a desired condition for complementariness. Case 5 shows that the maximum improvement given in Table 8 is for the candidate classifier which has the maximum dependence with the best classifier. Again a considerable improvement is possible for Case 4, where the output dependence between classifiers is quite low. A last observation on Table 8 is that the complementing classifier performance need not necessarily be very close to the performance of the best classifier for improvement to be possible. Again the maximum improvement is achieved by a complementing classifier with $p^e = 0.3$ while a much smaller improvement could be achieved with a much better performing classifier with $p^e = 0.21$.

8 Conclusion

This paper attempted to clarify the concepts of *output independence* and *complementariness* and their relations with the actual performance improvement achievable by optimal combination. The following have been the main contributions. Firstly, an Information Theoretic interpretation of a multiple classifier system is introduced and this enabled the use of measures from information theory to quantify relations between random variables representing events within such a system.

A measure for classifier output dependence is developed under this framework and it is shown that output independence plays no exclusive role in determining how much a classifier can complement another. A new concept called as *dominance of a classifier* is introduced to give a critical condition for performance improvement. Finally, another Information Theoretic measure is introduced to quantify the potential for improvement in such a system which have been supported by empirical justification. However, not all the questions raised within the scope of this paper could be answered and there exist several issues open for further research. The concept of *error independence* and its relation with performance improvement through combination remains an open issue. Also, the theoretical relation between the complementariness measure $\Delta I_{X_1 X_2}$ and the actual improvement remains to be established.

9 Acknowledgements

The authors would like to thank TÜBİTAK, The Turkish Scientific and Technical Research Council, for their financial support along the course of this research.

References

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

[2] Lei Xu, Adam Krzyżak, and Ching 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.

[3] Y.S. Huang and C.Y. 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.

[4] Tin Kam Ho. Recognition of handwritten digits by combining independent learning vector quantizations. In *Proceedings of 2nd International Conference on Document Analysis and Recognition*, pages 818–822, Tsukuba Science City, Japan, October 1993.

[5] Roberto Battiti and Anna Maria Colla. Democracy in neural nets: Voting schemes for classification. *Neural Networks*, 7(4):691–707, 1994.

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

[7] Josef Kittler, Mohamad Hatef, Rober P. W. Duin, and Jiri Matas. On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(3):226–239, March 1998.

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

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

[10] Younnès Bennani and Patrick Gallinari. Neural networks for discrimination and modelization of speakers. *Speech Communication*, 17:159–175, 1995.

[11] Mübeccel Demirekler and Afşar 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.

[12] Vlasta Radová and Joseph 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.

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

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

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

[16] Jon Atli Benediktsson, Johannes R. Sveinsson, Okan K. Ersoy, and Philip H. Swain. Parallel consensual neural networks. *IEEE Transactions on Neural Networks*, 8(1):54–64, January 1997.

[17] Jon Atli Benediktsson and Philip H. Swain. Consensus theoretic classification methods. *IEEE Transactions on Systems, Man and Cybernetics*, 22(4):688–704, July/August 1992.

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

[19] James Shih-Jong Lee, Jenq-Neng Hwang, Daniel T. Davis, and Alan 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] Yu Hen Hu, Surekha Palreddy, and Willis 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.

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

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

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

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

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

[26] Kagan Tumer and Joydeep Gosh. Estimating the bayes error rate through classifier combining. In *Proceedings of IAPR International Conference on Pattern Recognition*, pages 695–699, Vienna, Austria, 1996.

[27] Afşar Saranlı and Mübeccel Demirekler. A statistical unified framework for rank-based multiple classifier decision combination. To appear in *Pattern Recognition*.

[28] Afşar Saranlı and Mübeccel Demirekler. Rank-based multiple classifier decision combination: A theoretical study. In *Proceedings of the IEEE International Workshop on Intelligent Signal Processing*, pages 51–56, Budapest, Hungary, September 1999.

[29] Afşar Saranlı. *A Unifying Theory for Rank-Based Multiple Classifier Systems with Applications in Speaker Identification and Speech Recognition*. PhD thesis, Dept. of Electrical and Electronics Engineering, Middle East Technical University, Ankara, Turkey, January 2000.

[30] Robert J. McEliece. *The Theory of Information and Coding*, volume 3 of *Encyclopedia of Mathematics and Its Applications*. Addison-Wesley Publishing Co., London, 1977.