

Dependency Relationship Based Decision Combination in Multiple Classifier Systems*

Hee-Joong Kang¹ and Jin H. Kim

Computer Science Department and Center for Artificial Intelligence Research
Korea Advanced Institute of Science and Technology
373-1 Kusong-dong Yusong-gu Taejon 305-701, Korea

Abstract

Although many decision combination methods have been proposed, most of them did not focus on dependency relationship among classifiers. In combining multiple decisions, that makes classification performance of combining multiple decisions be degraded and biased, in case of adding highly dependent inferior classifiers. To overcome such weaknesses and obtain robust classification performance, the present study used dependency relationship for better combining multiple decisions. In order to identify dependency relationship by observing outputs of multiple classifiers, two methods are used on the basis of first-order dependency relationship. One is to use the concept of mutual information, and the other one is to use the concept of statistically measured association. The first-order dependencies identified are used to combine multiple decisions, using Bayesian formalism. A number of multiple classifier systems are applied to totally unconstrained on-line handwritten numerals and the English alphabet recognition. The experimental results show that the classification performance of a multiple classifier system is superior to that of individual classifiers. Also, they show that considering the dependency relationship outperforms others in accuracy, when the highly dependent inferior classifiers are added.

1 Introduction

Generally, two directions have been developed in the area of character and pattern recognition for improving classification performance. One is to improve the classification performance of a classifier itself. The other one is to improve a multiple classifier system which consists of a set of classifiers and a decision combination method. Although a number of classifiers are available, none of them is as good as expected. The major difficulty comes

¹This work is partly supported by the Notepad Consortium and the Offline Consortium.

*To whom all the correspondence should be addressed. He also works for Samsung Electronics, Co.

from the fact that although a number of features in diversified forms are available, but it is not easy to lump them together into a single classifier. The multiple classifier system is motivated from the assumption that a number of complementary features or classification algorithms can facilitate the classification performance, if they are used simultaneously. This research idea in combining multiple decisions seems to be promising.

Many studies in multiple classifier systems have focused on the fact that a classifier corresponds to an expert in multiple experts group. They conducted a study about the methodology of integrating multiple decisions from different classifiers [Mandler and Schuermann, 1988, Hull *et al.*, 1990, Suen *et al.*, 1990, Xu *et al.*, 1992, Franke and Mandler, 1992, Huang and Suen, 1993, Ho 1992, Ho *et al.* 1994]. Although many studies have been conducted for better combining multiple decisions, most of them did not focus on dependency relationship among classifiers. Instead, they considered multiple classifiers as being independent.

As seen in previous studies, if the highly dependent inferior classifiers are added to a multiple classifier system, and they have equal weights, then the classification performance of it can be degraded and biased by some of those decision combination methods. The motivation of this paper is to overcome the shortage of previous studies.

It is desirable to construct a multiple classifier system in which classifiers complement each other for better performance, and to combine multiple decisions by using the dependency relationship, if necessary. In this paper, we will provide some methods in identifying the dependency relationship, and in combining multiple decisions, using Bayesian formalism.

To identify dependency relationship of a classifier from A classifiers, we should consider all the possible cases of lower order subdistributions from the definition of product approximation by Lewis [Lewis, 1959]. Because that requires enormous computations and storage spaces, we will approximate K -dimensional distribution with the $K - 1$ first-order dependency distributions by considering only the classification results of classifiers, and by using two identification methods proposed. Two methods are considered for the work. One of them depends on the principle of maximizing mutual information by Chow [Chow and Liu, 1968], and the other one is to use

maximally measured association. The first-order dependencies identified are used in order to combine multiple decisions, using Bayesian formalism.

All the classifiers were created by Hidden Markov Models (HMMs). The HMMs have been used as a framework of on-line cursive script recognition [Sin *et al.*, 1994] and can model well for both variations of temporal sequences and spatial movements. Also, they were used as components of a multiple classifier system. A number of multiple classifier systems are applied to totally unconstrained on-line numerals and the English alphabet recognition.

In our experiment, we will apply a number of decision combination methods to combine multiple decisions in multiple classifier systems, and to compare their classification performances each other. In particular, several combinations of multiple classifiers for demonstrating the effects of highly dependent classifiers will be carried out, and then tested by decision combination methods proposed.

2 Related Works

A number of studies related to the idea of using multiple classifiers for improving classification performance will be described. The researches of combining multiple decisions are divided according to the types of classification results. These types include measurement scores of classes, rankings of classes, and single top choice of classes. From the measurement scores of classes, we can assign a ranking to each class by ordering classes as to their scores. Single top choice of a class is chosen by the best measurement score or by the first ranked.

Decision combination methods based on the measurement scores are averaged Bayes classified

[Xu *et al.*, 1992]

and an integration of multiple classification results using fuzzy integral [Tanahi and Keller, 1990] or fuzzy logic. In case of single top choice, there are voting methods [Suen *et al.*, 1990, Franke and Mandler, 1992, Xu *et al.*, 1992], the use of Bayesian formalism under an independence assumption [Xu *et al.*, 1992], the use of Dempster-Shafer formalism used in evidential reasoning [Mandler and Schuermann, 1988, Franke and Mandler, 1992, Xu *et al.*, 1992], and Behavior-Knowledge Space (BKS) [Huang and Suen, 1993]. In particular, in the use of Bayesian formalism by "

Xu *et al.* [Xu *et al.*,

1992], they assume that the classifiers are independent because they use independent feature sets or they are trained independently. This approach happens to cause a problem when they are not independent. Ho *et al.* [Ho, 1992, Ho *et al.*, 1994] support the use of rankings and have studied the decision combination methods using such rankings from multiple classifiers which include the highest rank method, Borda count method, and logistic regression. Borda count method is well known one of social choice functions used in our experiment. In this paper, we concentrate on the type of ranking decisions due to the advantages of rankings.

Lewis [Lewis, 1959] tackled a problem of approximating a n -th-order binary distribution by a product of several of its component distributions of lower order using the idea of extension by Hartmarue. He showed that

the product approximation, under suitably restricted conditions, had the property of minimum information. The approximation method was based on an information measure for the closeness of two distributions and on the criterion of maximum entropy. Two or more proposed approximations could be compared and the best one be selected without any knowledge of actual distribution beyond that given by the approximations. In other words, the process of comparison consists of selecting that approximation containing the greatest amount of correlation. However, the problem of selecting a set of component distributions of a given complexity to compose the best approximation remained unsolved.

Chow *et al.* [Chow and Liu, 1968] studied to solve the unsolved problem by Lewis and to best approximate an n -th-order distribution by a product of $n - 1$ second-order component distributions. A method was presented to approximate optimally an n -dimensional discrete probability distribution by a product of second-order distributions, or the distribution of the first-order tree dependencies. To find an optimum set of $n - 1$ first-order dependency relationship among n variables, a procedure was derived to yield an approximation of a minimum difference in information. An optimal procedure was on the basis of maximizing the mutual information between two variables and Maximum Weight Spanning Tree (MWST) algorithm by Kruskal.

3 Multiple Classifier Systems

A Multiple Classifier System (MCS) consists of a set of classifiers and a decision combination method [Ho, 1992]. In this paper, a set of classifiers is built based on a HMM structure and a few stochastic modeling methods [Sin *et al.*, 1994, Sin and Kim, 1994]. The HMM structure is left-to-right transitional and consists of 8 state nodes and only left-to-right arcs. In order to create classifiers consisting of HMMs modeled from training data, we use a few stochastic modeling methods which are standard modeling (i.e. Baum-Welch algorithm), duration modeling, and nonstationary modeling. One HMM was modeled to represent one class at training. Although they have the same HMM structure, they would be different classifiers if they are modeled by different modeling methods.

The graphical representations of stochastic modeling methods are shown in Figure 1 by focusing on transition probabilities at state nodes. Standard modeling method is easy and takes a short time to train the HMMs, but it does not properly model duration information of a pattern. Duration modeling method supplements such a weakness of standard modeling method, but it takes a long time to train. Nonstationary modeling method is proposed by Sin *et al.* [Sin and Kim, 1994] for overcoming the weakness of duration modeling method which does not properly consider the duration information. Their approach is based on the idea that the duration in a state node should be modeled as a function of duration period, and by that idea it can properly consider the duration information. This method has an advantage of the best modeling for typical patterns, but it takes a very long time to train and it has a disadvantage of the worst

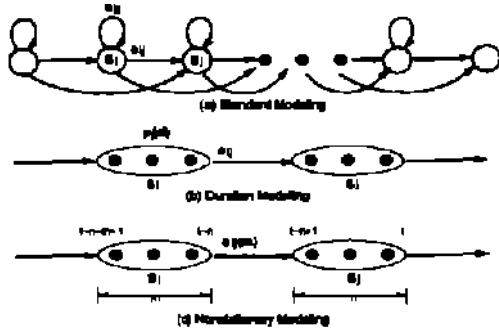


Figure 1 Stochastic modeling methods of HMMs

modeling for unusual patterns. A MCS is a good choice why there are no complete stochastic modeling methods for improving the classification performance.

For the recognition of numerals and the English alphabet characters, a classifier consists of 10 HMMs, or 26 HMMs, or 26 HMMs which are stochastically modeled for respective domain problem and assigns a likelihood score as to degree of match to each HMM for a given input. So, we have to convert the likelihood scores of HMMs into the rankings of classes modeled to HMMs before combining multiple decisions.

In this paper, a number of MCSs are constructed. One of them consists of three classifiers shown in Figure 2 which are based on the HMMs trained by three stochastic modeling methods. It is called by *Base-type* in our experiment. And the others are constructed by adding

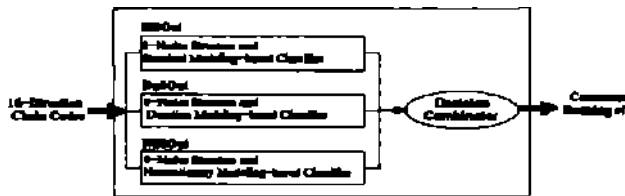


Figure 2 A MCS consists of three classifiers

the highly dependent inferior or superior classifier to the previous three classifiers which is respectively denoted by *St80Out*, *Du80Out* and *NS80Out*. The highly dependent classifier is made by faking one of the previous three classifiers.

Multiple decisions from individual classifiers of MCS will be combined according to a combining rule assigned to *Decision Combinator* and a winner class will be decided or not. The input of all classifiers are 16-directional chain codes in common. But different stochastic modeling methods cause classifiers to classify differently for a given input. The output of each classifier is converted as the rankings of classes, and then they are used as the input of *Decision Combinator*. Any classifier producing the rankings of classes can join into our MCS.

Only classifiers producing valid decisions take part in a process of *Decision Combinator*. Also, only classes having valid rankings are used in decision combination. Under those constraints, *Decision Combinator* includes a number of decision combination methods available in

this paper. They are voting methods, some of social choice functions, BKS method, the uses of Bayesian formalism under an independence assumption and dependency relationship among classifiers.

4 Dependency Relationship Based Decision Combination

On the assumption that multiple classifiers are independent, the use of Bayesian formalism is proposed by Xu et al [Xu et al, 1992]. In this paper, since an independence assumption would not be appropriate if some highly dependent classifiers join in multiple classifiers, we focus on the dependency relationship among classifiers. We will propose some methods for identifying the dependency relationship and combining multiple decisions, using Bayesian formalism. We will denote a set of classes by $\{M_1, \dots, M_M\}$, a set of classifiers by $C = \{C_1, \dots, C_K\}$, and an input pattern by z .

4.1 Identifying the Dependency Relationship

In order to build a K -dimensional joint probability distribution in a direct way, we can use K -dimensional samples observed. But when the K becomes larger and each classifier takes one of a set of M distinct classes, it requires much computation and massive storage of the order of M^K , and that is impractical in theoretical viewpoints. In other way, which is based on the *chain rule* of probability, we can express the joint probability distribution in terms of conditional probabilities.

$$P(C_1, \dots, C_K) = P(C_1)P(C_2|C_1) \dots P(C_K|C_1, \dots, C_{K-1})$$

In considering the dependency relationship, we assume that it is determined by the observations of outputs of individual classifiers given the same inputs. With A classifiers, the precise dependency relationship should be determined by considering maximum $K-1$ order dependencies by the definition of *chain rule*, since we do not know in advance the exact order of dependency relationship among them. When K classifiers are applied to the same input x , K events, $C_k(x) = M_j, k = 1, \dots, K$ and $j = 1, \dots, M$, will happen. Basing on the fact that it is desirable to consider first-order dependency relationship rather than to assume independence of classifiers, we will approximate the joint probability distribution with most appropriate second-order conditional probabilities based on the first-order dependency relationship.

$$P(C_1, \dots, C_K) = P(C_1)P(C_2|C_{i(1)}) \dots P(C_K|C_{i(K)})$$

Then the problem is how to identify an optimal set of first-order dependencies from classification results of classifiers. Our approaches to tackle that problem are to use two methods. One is to maximize the mutual information and the other one is to maximize the measures of association between classifiers.

Mutual Information

When we approximate K -dimensional distribution with a set of $K-1$ first- and second-order component distributions, we should identify the first-order dependence distribution such that the relative entropy, $I(P(C), P_i(C))$,

of unknown true distribution $P(C)$ and the first-order dependence distribution $P_i(C)$ a set of all possible first-order dependence trees is as small as possible [Lewis 1959, Chow and Liu, 1968] When the relative entropy is zero, we assure that a couple of distributions involved are equal

$$P_i(C) = \prod_{j=1}^K P(C_j | C_{i(j)})$$

From the expression of $I(P(C), P_i(C))$, minimizing the relative entropy is equivalent to maximizing the sum of mutual information (i.e. $\sum_{j=1}^K M(C_j, C_{i(j)})$) between a classifier C_j and a classifier $C_{i(j)}$. A classifier $C_{i(j)}$ is the parent of a classifier C_j in first-order dependency relationship. A root node in a dependence tree does not have a parent node. A $H(C)$ means an entropy of distribution C and C_j is a component of C . The definition of mutual information is described in [Chow and Liu, 1968, Gallager, 1968] and the average mutual information by Gallager means the mutual information by Chow et al which is used in this paper. To obtain the maximum sum of mutual information over all classifiers, we assign each classifier to a node of a possible dependence tree and assign the quantity of mutual information to a branch weight of the dependence tree. And then, we compute a maximum weight spanning tree over all the possible dependence trees and finally identify an optimal set of first-order dependencies from that tree.

$$\begin{aligned} I(P(C), P_i(C)) &= \sum_C P(C) \log \frac{P(C)}{P_i(C)} \\ &= \sum_C P(C) \log P(C) - \sum_C P(C) \sum_{j=1}^K \log P(C_j | C_{i(j)}) \\ &= \sum_C P(C) \log P(C) - \sum_{j=1}^K \log P(C_j) - \sum_C P(C) \sum_{j=1}^K \log \frac{P(C_j, C_{i(j)})}{P(C_j)P(C_{i(j)})} \\ &= -H(C) + \sum_{j=1}^K H(C_j) - \sum_{j=1}^K M(C_j, C_{i(j)}) \\ H(C) &= -\sum_C P(C) \log P(C) \\ P(C_j, C_{i(j)}) &= \sum_{C_j, C_{i(j)}} P(C_j, C_{i(j)}) \log \frac{P(C_j, C_{i(j)})}{P(C_j)P(C_{i(j)})} \end{aligned}$$

Measures of Association

Whole the mutual information contains the average amount of uncertainty to be resolved between two distributions, the measures of association measure the strength of association between two distributions from sampling data. For applying some coincidence measures such as correlation coefficient in numerical data to qualitative nominal data classes, we will use some measures of

association as criteria of statistical dependence. While the correlation coefficient represents the degree of linear dependence between numeric data, our classifiers output non-numerical nominal data classes. Therefore, the correlation coefficient is not appropriate to identify the dependency relationship of our classifiers.

In computing measures of association in statistics, there are Cramer's Value, i.e. V , and Contingency Coefficient, i.e. CC , computed from Pearson χ^2 statistic, Entropy symmetric measure, i.e. E_{sym} , from reduction in uncertainties in predicting the relationship between two classifiers, and Lambda symmetric measure, i.e. λ_{sym} , as an index of predictive association [Hays and Winkler, 1971]. They are non-negative real values. We think that those measures represent the degree of dependence relationship between classifiers. For expressing those measures, let N be the number of input samples, I be the number of output classes by classifier i , J be the number of output classes by classifier j , O_{ij} or n_{ij} be the number of observed outcomes from classifiers i and j , n_{i+} be $\sum_{j=1}^J n_{ij}$, n_{+j} be $\sum_{i=1}^I n_{ij}$, and E_{ij} be the number of expected outcomes from classifiers i and j , that is $\frac{n_{i+}n_{+j}}{N}$.

$$\begin{aligned} \chi^2 &= \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \\ V &= \sqrt{\frac{\chi^2}{N \min(I-1, J-1)}} \\ CC &= \sqrt{\frac{\chi^2}{\chi^2 + N}} \\ H(i) &= -\sum_{j=1}^J \frac{n_{i+}}{N} \log \frac{n_{i+}}{N} \\ H(j) &= -\sum_{i=1}^I \frac{n_{+j}}{N} \log \frac{n_{+j}}{N} \\ H(i,j) &= -\sum_{i=1}^I \sum_{j=1}^J \frac{n_{ij}}{N} \log \frac{n_{ij}}{N} \\ E_{sym} &= 2 \left[\frac{H(i) + H(j) - H(i,j)}{H(i) + H(j)} \right] \\ \lambda_{sym} &= \frac{(\sum_i \max_j n_{ij} + \sum_j \max_i n_{ij} - \max_{i,j} n_{ij} - \max_{i,j} n_{i+})}{(2N - \max_{i,j} n_{ij} - \max_{i,j} n_{i+})} \end{aligned}$$

We can assign the values of those measures to branch weights of possible dependence tree, as in case of mutual information. To identify an optimal set of first-order dependencies using one of the measures of association is to maximize the total sum of values of measures of association used using MWST algorithm and to select a dependence tree having such maximum sum from all the possible dependence trees. Although those measures have different scales, they coincide in an optimal set of first-order dependencies in our experiment.

4.2 The use of Bayesian Formalism

With an optimal set of first-order dependencies, single confusion matrices, and pairwise joint con-

fusion matrices, we can compute the group confusion probabilities of each output class to every class at test stage. For each class M_i , using Bayesian theorem and an optimal set of first-order dependencies, $P(C) = \prod_{j=1}^K P(C_j|C_{i(j)})$, have we

$$\begin{aligned} Bel(M_i) &= P(x \in M_i | C_1(x) = M_1, \dots, C_K(x) = M_K) \\ &= P(x \in M_i) \frac{P(C_1(x) = M_1, \dots, C_K(x) = M_K | x \in M_i)}{P(C_1(x) = M_1, \dots, C_K(x) = M_K)} \\ &= P(x \in M_i) \frac{\prod_{j=1}^K P(C_j(x) = M_j | C_{i(j)}(x) = M_{i(j)}, x \in M_i)}{\prod_{j=1}^K P(C_j(x) = M_j | C_{i(j)}(x) = M_{i(j)})} \\ &= P(x \in M_i) \frac{\prod_{j=1}^K P(x \in M_i | C_j(x) = M_j, C_{i(j)}(x) = M_{i(j)})}{\prod_{j=1}^K P(x \in M_i | C_{i(j)}(x) = M_{i(j)})} \end{aligned}$$

From the above expression, if we remove constant terms,

$$Bel(M_i) = \eta \frac{\prod_{j=1}^K P(x \in M_i | C_j(x) = M_j, C_{i(j)}(x) = M_{i(j)})}{\prod_{j=1}^K P(x \in M_i | C_{i(j)}(x) = M_{i(j)})}$$

with η as a constant that ensures that $\sum_{i=1}^M Bel(M_i) = 1$. If a classifier $C_{i(j)}$ is null, a classifier C_j is the root node of dependence tree. In this case, $P(x \in M_i | C_j(x) = M_j, C_{i(j)}(x) = M_{i(j)})$ is approximated by $P(x \in M_i | C_j = M_j)$. Finally, depending on those $Bel(M_i)$ values, we can classify x into a class according to the decision rule $E(x)$ given below

$$E(x) = \begin{cases} i, & \text{if } Bel(x) = \max_{M_i \in M} Bel(M_i) \\ M + 1, & \text{otherwise} \end{cases}$$

5 Analysis of Some Empirical Results

An *Base-type* Multiple Classifier System (MCS) consists of only original three classifiers (see Figure 2). For demonstrating the effects of highly dependent classifiers to combining multiple decisions, we have built a number of MCSs by adding the highly dependent classifier to the *Base-type* MCS by faking one of its component classifiers. For example, "*St8Out faked*" MCS consists of original three classifiers and the one created by faking *St8Out* classifier. We apply them to totally unconstrained on-line numerals and the English alphabet recognition.

Besides voting methods with absolute majority principle and simple majority principle, some of social choice functions are implemented [Hwang and Lin, 1987]. Social choice functions include Condorcet function, Borda function, and Nanson function (also called by Borda elimination method with the lowest Borda score). These functions have been widely used to choose the winner based on group consensus from ranking decisions of alternatives in the area of group decision support systems.

No training stage is required for applying voting methods and some of social choice functions, but it is needed for applying BKS method and some methods in Bayesian formalism. With training data, we use 4088 items of numerals written by 13 writers, 3749 items of the English lowercases written by 19 writers, and 2464 items of the English uppercases written by 19 writers. And with

test data, there are 988 items of numerals written by 10 writers, 1684 items of the English lowercases written by 9 writers, and 1169 items of the English uppercases written by 9 writers. The writers of the test data are different from those of the training data. The sign words in following tables are denoted in Section 3. The sign word $u^{?n}$ means a rejection and contains the number of samples rejected. The recognition rates are computed in percentage from the rank of the winner matched with the input label class. The recognition rates of classifiers for test data in application problems are shown in Table 1.

Classifier	Numerals		Lowercase		Uppercase	
	1st	?	1st	?	1st	?
St8Out	93.09	6	78.62	21	88.37	21
Du8Out	92.28	1	82.30	9	91.02	6
NS8Out	94.36	13	86.06	46	87.61	60

Table 1 Recognition rates of classifiers for test data

Comb Method	Base-type		Du8Out faked		NS8Out faked	
	1st	?	1st	?	1st	?
Voting (Maj)	94.99	7	91.98	62	94.29	19
Voting (Abs)	94.99	7	91.98	63	94.29	19
Condorcet Fn	94.99	0	91.98	0	94.29	0
Borda Fn	95.09	1	91.98	1	94.59	1
Nanson Fn	94.99	1	91.98	1	94.29	1
Indep. Bayes	95.29	1	95.49	1	95.19	1
Mut. Inf. Bay	95.39	1	95.59	1	95.39	1
V. Bayes	95.39	1	95.59	1	95.39	1
CC Bayes	95.39	1	95.59	1	95.39	1
E _{sym} Bayes	95.39	1	95.59	1	95.39	1
λ _{sym} Bayes	95.39	1	95.59	1	95.39	1
BKS Method	93.99	30	93.99	30	93.99	30

Table 2 Recognition rates of MCSs for numerals

From the experimental resultB of test numerals (see Table 1 and Table 2), the recognition rates of voting methods and social choice functions are lowered when inferior *Du8Out* or superior *NS8Out* classifiers are faked and added. The extent of degradation in recognition rates is larger when an inferior classifier is faked and added. There is hardly difference in classification performance among some methods of Bayesian formalism. Although one of classifiers in *Base-type* MCS is faked and added, the recognition rates of BKS method are unchanged. The recognition rates of dependency relationship based decision combination methods show a little higher than those of an independence assumption based decision combination in use of Bayesian formalism, but that is not statistically significant by *t*-test at significance level 0.1.

From the experimental results of test lowercases (see Table 1 and Table 3), the recognition rates of voting methods and social choice functions are lowered when inferior *St8Out* or superior *NS8Out* classifiers are faked and added. The extent of degradation in recognition rates is much larger when an inferior classifier is faked and added. The recognition rates of dependency relationship based decision combination methods show higher than those of an independence assumption based decision combination in use of Bayesian formalism by 0.5 - 1.4%. The unproved correctness of a *Base-type* MCS by

Comb Method	Base-type		Si8Out faked		NS8Out faked	
	1st	?	1st	?	1st	?
Voting (Maj)	84 80	64	78 86	149	84 62	78
Voting (Abs)	84 80	64	78 62	192	84 14	118
Condorcet Fn	85 21	0	78 74	0	84 66	0
Borda Fn	85 27	7	82 60	7	85 10	7
Nanson Fn	85 33	7	80 70	7	84 68	7
Indep Bayes	87 00	7	86 34	7	87 23	7
Mut Inf Bay	87 77	7	87 77	7	87 77	7
V Bayes	87 77	7	87 77	7	87 77	7
CC Bayes	87 77	7	87 77	7	87 77	7
E _{sym} Bayes	87 77	7	87 77	7	87 77	7
λ _{sym} Bayes	87 77	7	87 77	7	87 77	7
BKS Method	85 87	134	85 87	134	85 87	134

Table 3 Recognition rates of MCSs for the lowercases

the dependency relationship is statistically significant by (-test at significance level 0.05, and in case of a *Si8Out faked* MCS, its improved correctness is significant by *t*-test at significance level 0.005, and in case of a *NS8Out* MCS, its improved correctness is significant by (-test at significance level 0.1. We will describe one example to perform (-test in a *St8Out faked* MCS to verify whether the improved correctness by considering the dependency relationship is statistically significant or not. Like the case of numerals, the recognition rates of BKS method are unchanged.

- H_0 The average correctness of dependency relationship based decision combination (i.e. Mut Inf Bay) is equal or less than that of an independence assumption based decision combination (i.e. Indep Bayes) (i.e. $\mu_D \leq 0$ and D is difference)
- H_a An alternative to H_0 (i.e. $\mu_D > 0$)
- Let n be the number of writers,

$$T = \frac{\bar{D} - 0}{S_D / \sqrt{n}} = 3.51180$$

where

$$\bar{D} = \frac{1}{n} \sum_{i=1}^n D_i = \frac{1}{9}(13.16) = 1.46222$$

and

$$S_D^2 = \frac{1}{n-1} \sum_{i=1}^n (D_i - \bar{D})^2 = 1.56031$$

- We can reject H_0 with 8 degrees of freedom because of $T = 3.51180 > t_{0.005} = 3.355$. It means that the average correctness of dependency relationship based decision combination is improved with 99.5% of belief interval.

From the experimental results of test uppercases (see Table 1 and Table 4), the recognition rates of voting methods and social choice functions are a little lowered when inferior *NS8Out* or superior *Du8Out* classifiers are faked and added. The extent of degradation in recognition rates is larger when an inferior classifier is faked

Comb Method	Base-type		NS8Out faked		Du8Out faked	
	1st	?	1st	?	1st	?
Voting (Maj)	90 33	32	88 02	32	89 39	50
Voting (Abs)	90 33	32	87 86	66	88 79	63
Condorcet Fn	90 69	0	87 94	0	88 96	0
Borda Fn	90 08	5	89 14	6	89 14	5
Nanson Fn	90 33	5	88 11	5	89 39	5
Indep Bayes	90 69	5	89 48	5	91 27	5
Mut Inf Bay	90 69	5	90 68	5	90 69	5
V Bayes	90 69	5	90 68	5	90 69	5
CC Bayes	90 69	5	90 68	5	90 69	5
E _{sym} Bayes	90 69	5	90 68	5	90 69	5
λ _{sym} Bayes	90 69	5	90 68	5	90 69	5
BKS Method	88 71	88	88 71	88	88 71	88

Table 4 Recognition rates of MCSs for the uppercases

and added. The improved correctness of a *NS8Out faked* MCS by the dependency relationship is statistically significant by (-test at significance level 0.05, and in case of a *Du8Out faked* MCS, its degraded correctness is in significant- by (-test at significance level 0.1. Like the case of numerals, the recognition rates of BKS method are unchanged.

They are the straightforward instances of problems before mentioned. Therefore, it is problematic to add simply any classifiers to existing multiple classifier system and to apply voting methods or some of social choice functions. The classification performance of BKS method is also lowered by adding classifiers.

From the analysis of some empirical results, we come to some conclusions.

- The classification performance of a .Base-type MCS is almost superior to that of individual classifiers.
- Voting methods and social choice functions show similar behaviors in combining multiple decisions.
- All the combination methods in Bayesian formalism almost outperforms the other combination methods.
- Incorporating the dependency relationship into Bayesian formalism helps improving the classification performance of a MCS, especially when the highly dependent inferior classifiers are added to it. Some empirical results for lowercase significantly support our assertion by showing statistically significant (-test results).
- It is problematic to add simply any classifiers or to combine multiple decisions without considering the dependency relationship among classifiers. Because the highly dependent inferior classifiers can degrade the classification performance of a MCS, if voting methods or some of social choice functions are used.
- The classification performance of BKS method is unchanged when one of component classifiers are faked and added.

6 Conclusion and Further Works

In order to prove the effectiveness of dependency relationship for combination of multiple decisions, and for construction of multiple classifier systems, we proposed a couple of methods for identifying an optimal set of first-order dependencies approximated from sampling data,

and a decision combination method of multiple decisions in Bayesian formalism, using the first-order dependencies identified. This research applied the proposed methods to totally unconstrained on-line numerals and the English alphabet recognition. The results suggest that the dependency relationship should be considered not only for combining multiple decisions, but also for constructing multiple classifier systems.

Further studies should examine both the robust criteria for identifying the dependency relationship correctly, and the good approximation method(s) for obtaining the dependency relationship of a group of classifiers with the high-order dependency relationship. These proposed ideas will play important roles in resolving the following questions: how many classifiers are chosen⁷, what kind of classifiers should be used⁷, how a subset of classifiers is selected dynamically⁷, and how multiple decisions are combined⁷.

Acknowledgments

The authors wish to thank Bongkee Sin for providing the classification results of three individual classifiers St8Ovt, Du80ut, and NSSOut which form the basis of the experiments performed. The authors would also like to thank the anonymous reviewers for their helpful comments.

References

- [Chow and Liu, 1968] C K Chow and C N Liu "Approximating Discrete Probability Distributions with Dependence Trees" *IEEE Transactions on Information Technology*, 14(3) 462-467, 1968
- [Franke and Mandler, 1992] Jurgen Franke and Eberhard Mandler "A Comparison of Two Approaches for Combining the Votes of Cooperating Classifiers" In *Proceedings of the 11th IAPR International Conference on Pattern Recognition*, volume 2, pages 611-614, 1992
- [Gallager, 1968] Robert G. Gallager *Information Theory and Reliable Communication* John Wiley and Sons, Inc., 1968
- [Hays and Winkler, 1971] William L Hays and Robert L Winkler *Statistics probability, inference, and decision* Holt, Rinehart and Winston, Inc., 1971
- [Ho et al., 1994] Tin Kam Ho, Jonathan J Hull, and Sargur N Srihari "Decision Combination of Multiple Classifier Systems" *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(1) 66-75, 1994
- [Ho, 1992] Tin Kam Ho "A Theory of Multiple Classifier Systems and Its Application to Visual Word Recognition" PhD thesis, Dept of Computer Science, SUNY at Buffalo, 1992
- [Huang and Suen, 1993] Y S Huang and C Y Suen "An Optimal Method of Combining Multiple Classifiers for Unconstrained Handwritten Numeral Recognition" In *Proceedings of the 3rd International Workshop on Frontiers in Handwriting Recognition*, pages 11-20, 1993
- [Hull et al., 1990] Jonathan J Hull, Alan Comiskey, and Tin-Kam Ho "Multiple Algorithm for Handwritten Character Recognition" In *Proceedings of the 1st International Workshop on Frontiers in Handwriting Recognition*, pages 117-129, 1990
- [Hwang and Lin, 1987] Ching-Lai Hwang and Ming-Jeng Lin *Group Decision Making under Multiple Criteria* Lecture Notes in Economics and Mathematical Systems Springer-Verlag, 1987
- [Lewis, 1959] P M Lewis "Approximating Probability Distributions to Reduce Storage Requirement" *Information and Control*, 2 214-225, 1959
- [Mandler and Schuermann, 1988] Eberhard Mandler and J Schuermann "Combining the classification results of independent classifiers based on the Dempster-Shafer theory of evidence" In E S Gelsema and L N Kanal, editors, *Uncertainty in Artificial Intelligence*, pages 381-393 Elsevier Science Publishers, 1988
- [Sin and Kim, 1994] Bongkee Sin and Jin H Kim "Nonstationary Hidden Markov Model" Submitted to *Signal Processing*, 1994
- [Sin et al., 1994] Bongkee Sin, Jin-Yong Ha, Se-Chang Oh, and Jin H Kim "Network-Based Approach to On-line Cursive Script Recognition" Submitted to *IEEE Transactions on Systems, Man, and Cybernetics*, 1994
- [Suen et al., 1990] C Y Suen, C Nadal, T A Mai, R Legault, and L Lam "Recognition of Totally Unconstrained Handwritten Numerals Based on the Concept of Multiple Experts" In *Proceedings of the 1st International Workshop on Frontiers in Handwriting Recognition*, pages 131-143, 1990
- [Tanahi and Keller, 1990] Hossein Tanahi and James M Keller "Information Fusion in Computer Vision Using the Fuzzy Integral" *IEEE Transactions on Systems, Man, and Cybernetics*, 20(3) 733-741, 1990
- [Xu et al., 1992] Lei Xu, Adam Krzyzak, 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, 1992