

Product Approximation by Minimizing the Upper Bound of Bayes Error Rate for Bayesian Combination of Classifiers*

Hee-Joong Kang[†]
Hansung University
Division of Computer Engineering
Samsun-dong 3-ga, Sungbuk-gu, Seoul, Korea
hjkang@hansung.ac.kr

David Doermann
University of Maryland
Institute for Advanced Computer Studies
College Park, MD 20742-3275, USA
doermann@umiacs.umd.edu

Abstract

In combining multiple classifiers using a Bayesian formalism, a high dimensional probability distribution is composed of a class and decisions of classifiers. In order to do product approximation of the probability distribution, the upper bound of Bayes error rate, bounded by the conditional entropy of a class and decisions, should be minimized. A second-order dependency-based product approximation is proposed in this paper by considering the second-order dependency between the class and decisions. The proposed method is evaluated by combining the classifiers recognizing unconstrained handwritten numerals.

1. Introduction

In order to combine the decisions of multiple classifiers using a Bayesian formalism, a label class and decisions are represented in terms of a high dimensional probability distribution in the training stage. On the assumption that the decisions are conditionally independent of the given class, the high dimensional probability distribution is approximated with a product of two-dimensional component distributions and the decisions can be combined (Xu et al. [12]). This assumption can be regarded as the special case of the first-order dependency among components. The first-order dependency-based product approximation (DBPA) is proposed by Chow and Liu [2]. Later, Kang et al. proposed the second-order DBPA scheme in [6] and the third-order DBPA scheme in [5] by considering more than the first-order dependency among components in approximating the

probability distribution. These DBPAs do not have any constraint in dealing with the components to approximate the high dimensional probability distribution.

Another first-order DBPA is proposed by Wang and Wong [11] who define the class-patterns (CP) mutual information for considering the dependency between a class and patterns for product approximation. This is the main difference with the product approximations in [2]. Kang and Lee also applied the concept of CP mutual information to class-decisions (CD) relationship in combining multiple classifiers and tried to combine multiple classifiers with the product approximation derived from CD mutual information using Bayesian formalism [7]. Without any product approximation, direct full dependency between a class and decisions is considered in the method of Behavior-Knowledge Space (BKS) in [4]. However, the BKS method has both the possibility of high rejection rates due to unseen decisions and the exponential complexity in directly storing and estimating the high dimensional probability distribution. In this paper, another second-order DBPA scheme is proposed as the extended work of the first-order DBPA by Wang and Wong, using the CD mutual information.

The proposed method is evaluated by combining the classifiers recognizing unconstrained handwritten numerals from Concordia University [10] and the University of California, Irvine (UCI) [1]. Six classifiers are combined at an abstract level, where these classifiers were developed by using the features or methodologies in [8, 9]. The Bayesian combination methods based on the presented DBPAs are introduced in the recognition experiments as is the BKS method.

This paper is organized as follows. Section 2 explains the CD mutual information and the second-order DBPA scheme. Bayesian combination using the proposed second-order DBPA is defined in Section 3. Experimental results for evaluating the proposed DBPA with Bayesian combination methods are provided in Section 4 and the concluding remarks are given in Section 5.

* The support of this research, under DOD contract MDA90402C0406 and National Science Foundation grant EIA0130422 is gratefully acknowledged.

† Dr. Kang is a visiting researcher at the LAMP laboratory, University of Maryland.

2. Class-Decisions (CD) mutual information

Hellman and Raviv proved an inequality expression between the Bayes error rate P_e and the conditional entropy $H(M|C)$ of a class M and variables C , (Eq. (1) from [3]). This paper regards the variables as decisions. The Bayes error rate P_e is upper bounded by the conditional entropy $H(M|C)$. Thus, the CD mutual information $U(M; C)$ is defined from the conditional entropy in Eq. (1) and measures the degree of dependence between the class M and the decisions C , as:

$$P_e \leq \frac{1}{2}H(M|C) = \frac{1}{2}(H(M) - U(M; C)) \quad (1)$$

$$U(M; C) = \sum_m \sum_c P(m, c) \log \frac{P(m, c)}{P(m)P(c)} \quad (2)$$

where $H(M)$ is the entropy. It is obvious that minimizing the upper bound of P_e leads to maximizing the CD mutual information $U(M; C)$, since $H(M)$ does not depend on C .

When K decisions, C_1, \dots, C_K , are combined by, a second-order DBPA is obtained by considering the second-order dependency among the decision components in the probability distribution. The approximating distribution of C is defined in terms of three-dimensional distributions:

$$P_a(C_1, \dots, C_K) = \prod_{j=1}^K P(C_{n_j} | C_{n_{i2(j)}}, C_{n_{i1(j)}}), \quad (3)$$

such that $(0 \leq i2(j) \leq i1(j) < j)$ holds, and the approximating distribution of C and M is defined in terms of four-dimensional distributions:

$$P_a(C_1, \dots, C_K, M) = \prod_{j=1}^K P(C_{n_j} | C_{n_{i2(j)}}, C_{n_{i1(j)}}, M), \quad (4)$$

$$P_a(C_1, \dots, C_K | M) = \frac{1}{P(M)} \prod_{j=1}^K P(C_{n_j} | C_{n_{i2(j)}}, C_{n_{i1(j)}}, M), \quad (5)$$

such that $(0 \leq i2(j) \leq i1(j) < j)$ holds and C_{n_j} is conditioned on $C_{n_{i2(j)}}$, $C_{n_{i1(j)}}$, and M , and where (n_1, \dots, n_K) is an unknown permutation of integers $(1, \dots, K)$ and C_0 is a null component. Thus $P(C_{n_j} | C_0, C_0, M)$ is equal to $P(C_{n_j}, M)$, and $P(C_{n_j} | C_0, C_{n_{i1(j)}}, M)$ is equal to $P(C_{n_j} | C_{n_{i1(j)}}, M)$, by definition. The second-order dependency makes the CD mutual information expanded like the following expressions by using the Eqs. (3)–(5) and dropping the subscript n of C :

$$\begin{aligned} U(M; C) &= \sum_m \sum_c P(c, m) \log \frac{P(c|m)}{P(c)} \\ &= \sum_{m,c} P(c, m) \log \left[\frac{1}{P(M)} \prod_{j=1}^K P(C_j | C_{i2(j)}, C_{i1(j)}, M) \right] \end{aligned}$$

$$\begin{aligned} &= - \sum_c P(c) \log \prod_{j=1}^K P(C_j | C_{i2(j)}, C_{i1(j)}) \\ &= - \sum_m P(m) \log P(m) + \sum_{j=1}^K \sum_{m,c} P(c, m) \log P(C_j | C_{i2(j)}, C_{i1(j)}, M) \\ &= - \sum_{j=1}^K \sum_c P(c) \log P(C_j | C_{i2(j)}, C_{i1(j)}) \\ &= H(M) + \sum_{j=1}^K [I(C_j; C_{i2(j)}, C_{i1(j)}, M) - I(C_j; C_{i2(j)}, C_{i1(j)})] \quad (6) \end{aligned}$$

$$H(M) = - \sum_m P(m) \log P(m) \quad (7)$$

$$\Delta I(C_j; C_{i2(j)}, C_{i1(j)}) = I(C_j; C_{i2(j)}, C_{i1(j)}, M) - I(C_j; C_{i2(j)}, C_{i1(j)}) \quad (8)$$

From the above derived Eq. (6), maximizing $U(M; C)$ leads to maximizing $\sum_{j=1}^K \Delta I(C_j; C_{i2(j)}, C_{i1(j)})$ which is the total sum of Δ second-order CD mutual information, since remaining term $H(M)$ is also irrespective of C . Then, the next step is finding an optimal product set by the second-order dependency from all the permissible product sets. Finding the optimal product set by the second-order dependency is to select the maximum sum of Δ second-order CD mutual information covering Δ first-order CD mutual information, as described in the following algorithm. From the optimal product set, we can determine the unknown permutation (n_1, \dots, n_K) and their two unknown conditioned permutations $(n_{i2(1)}, \dots, n_{i2(K)})$ and $(n_{i1(1)}, \dots, n_{i1(K)})$.

Input:

A set of $(K + 1)$ -dimensional samples of C and M .

Output:

An optimal product set by the second-order dependency as per the Δ second-order CD mutual information.

Method:

1. Estimate two-, three-, and four-dimensional marginal distributions from the samples.
2. Compute the weights $\Delta I(C_j; C_{i(j)})$ and $\Delta I(C_j; C_{i2(j)}, C_{i1(j)})$ for all pairs, and triplets of classifiers from the estimated marginal distributions.
3. Compute the maximum weight sum consisted of Δ first-order and Δ second-order CD mutual information and find its associated optimal product set, as the following statements:

maxTweight = 0;

for $n = 1$ **to** no. of Δ first-order CD mutual information **do**

Tweight = weight of the n -th $\Delta I(C_j; C_{i(j)})$;

while (no. of untraversed classifiers) > 0 **do**

choose one of untraversed classifiers and mark it traversed;

choose the largest permissible Δ second-order CD mutual information associated with the chosen classifier and one traversed classifier among all traversed classifiers;

Tweight += weight of the chosen $\Delta I(C_j; C_{i2(j)}, C_{i1(j)})$;

end

maxTweight = MAX(*maxTweight*, *Tweight*);

store *maxTweight* and its associated Δ first-order and Δ second-order CD mutual information;

end

obtain maximum *maxTweight* and its associated Δ first-order and Δ second-order CD mutual information;

By using the systematic approach for product approximation, the order of dependency considered can be easily extended to the d th-order under permissible computing resources. Considering the d th-order dependency makes the approximating distributions Eqs. (3)–(5) changed as to the

order of dependency d . An optimal product set by the d th-order dependency consists of one by Δ first-order CD mutual information, one by Δ second-order CD mutual information, ..., one by Δ $(d - 1)$ st-order CD mutual information, and multiple (i.e. $(K - d)$) component distributions by Δ d th-order CD mutual information.

3. Bayesian combination using product approximation

After an optimal product set by the second-order dependency is found and all unknown permutations for it are determined, Bayesian combination of K classifiers is derived from using the Bayesian formalism and the optimal product set. For a hypothesized class m , its supported belief function $Bel(m)$ is defined by the following expressions using the Eq. (4):

$$\begin{aligned} Bel(m) &= P(m \in M | C_1, \dots, C_K) = \frac{P(C_1, \dots, C_K, M)}{P(C_1, \dots, C_K)} \\ &= \frac{\prod_{j=1}^K P(C_{n_j} | C_{n_{i2(j)}}, C_{n_{i1(j)}}, M)}{P(C_1, \dots, C_K)} \\ &\approx \eta \prod_{j=1}^K P(C_{n_j} | C_{n_{i2(j)}}, C_{n_{i1(j)}}, M), \end{aligned} \quad (9)$$

with η as a constant that ensures that $\sum_{i=1}^L Bel(m_i) = 1$ and (n_1, \dots, n_K) is an unknown permutation of integers $(1, \dots, K)$ where L is the number of classes. Therefore, the combination of classifiers by the second-order dependency is to determine a hypothesized class m which maximizes the supported belief function $Bel(m)$ in the Eq. (9). Depending on the belief value $Bel(m)$, we can choose a maximized posterior probability $P^*(m \in M | C_1, \dots, C_K)$, and then a combined decision is determined or not, according to the decision rule $D(C)$ given below:

$$D(C) = \begin{cases} M_i, & \text{if } Bel(m_i) = \max_{m_j \in M} Bel(m_j) \\ L + 1, & \text{otherwise.} \end{cases}$$

4. Experimental results

Six classifiers, $E_1, E_2, E_3, E_4, E_5, E_6$, are used for the recognition experiments of the unconstrained handwritten numerals from Concordia University [10] and the University of California, Irvine (UCI) [1]. These classifiers were developed by using the features in [8, 9] or by using the structural knowledge of numerals, such as bounding, centroid, and the width of horizontal runs or strokes, at KAIST and Chonbuk National University. Some of them are back-propagation singular or modular neural networks and the others are rule-based modular recognizers.

Classifiers E_2 and E_3 are modular architecture and use directional distance distribution and mesh features respectively. Classifiers E_1 and E_6 are singular architecture and use pixel distance function and contour features respectively. Since classifiers E_4 and E_5 were trained by the structural knowledge obtained from Concordia numerals, they are not very good on the UCI numerals due to high rejection rates. The performance of individual classifiers is shown in Table 1 with recognition and reliability rates for respective test data sets T and *windep*.

Classifier	T of Concordia		<i>windep</i> of UCI	
	Recog.	Relia.	Recog.	Relia.
E_1	96.00	96.00	93.77	93.77
E_2	95.95	95.95	97.11	97.11
E_3	84.45	96.24	91.82	96.95
E_4	90.95	99.02	67.67	93.11
E_5	88.15	98.38	70.01	94.80
E_6	94.15	94.15	96.66	96.66

Table 1. Performance of individual classifiers

The classifiers were combined with the test data sets T , *windep*, using the following combination methods: the BKS method in [4], and the several Bayesian combination methods as noted in Table 2. Among the Bayesian combination methods, the CIAB method was proposed in [12], and the ODB1, CODB1, and ODB2 methods were proposed in [6], and the CODB2 and ODB3 methods were proposed in [5], and the DODB1 and DODB2 methods are proposed in this paper by using the Δ first-order and Δ second-order CD mutual information, respectively.

Notation	Full Term
CIAB	conditional independence assumption-based
ODB1	first-order dependency-based
CODB1	conditional 1st-order dependency-based
ODB2	2nd-order dependency-based
CODB2	conditional 2nd-order dependency-based
ODB3	3rd-order dependency-based
DODB1	Δ 1st-order dependency-based
DODB2	Δ 2nd-order dependency-based

Table 2. Bayesian combination methods

The first experiment is to combine five classifiers selected from the six candidates, so six groups were made from 5G1 to 5G6. The best recognition rate in each group in Figs. 1 and 2 was primarily obtained by the DODB1 or DODB2 method.

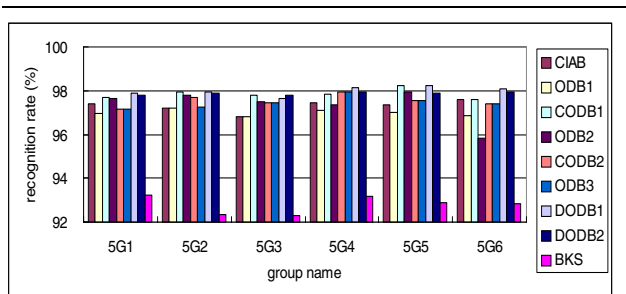


Figure 1. Results of five classifiers on *T*

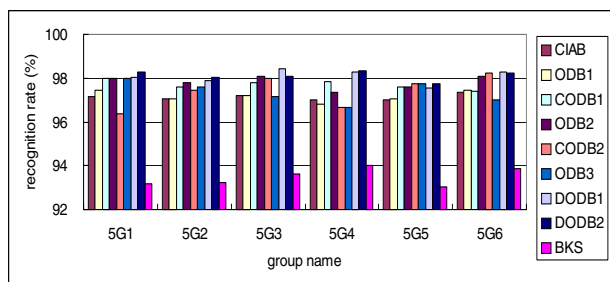


Figure 2. Results of five classifiers on *windep*

The second experiment is to combine all six classifiers, so two groups were made according to the source of test data. The best recognition rate in each group in Fig. 3 was obtained by the DODB1 or DODB2 method.

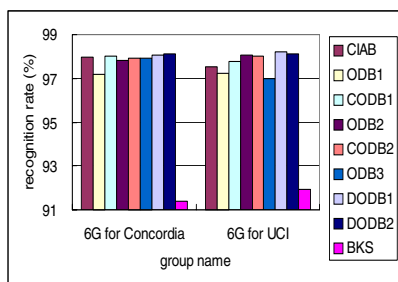


Figure 3. Results of six classifiers on *T* and *windep*

The experimental results supported that the proposed DBPA and its combination method contributed to improvement on the performance over other Bayesian combination methods by raising the class discrimination power with the CD mutual information, although it required larger storage needs than the previous Bayesian methods for computing the Δ *n*th-order CD mutual information. Particularly, the

low recognition rates of the BKS method might be caused by the lack of large enough and well representative training data sets.

5. Concluding Remarks

This paper extended the work of Wang and Wong to the second-order dependency and reviewed the dependency between a class and decisions with the defined CD mutual information and the BKS method. An algorithm using the Δ second-order CD mutual information for the second-order dependency was also described in this paper. In order to raise the class discrimination power in combining multiple classifiers, the upper bound of Bayes error rate should be minimized and thus the best recognition rates were obtained with the proposed DBPA.

References

- [1] C. Blake and C. Merz. UCI repository of machine learning databases, 1998.
- [2] C. K. Chow and C. N. Liu. Approximating Discrete Probability Distributions with Dependence Trees. *IEEE Trans. on Information Theory*, 14(3):462–467, 1968.
- [3] M. E. Hellman and J. Raviv. Probability of error, equivocation, and the Chernoff bound. *IEEE Trans. on Information Theory*, IT-16:368–372, 1970.
- [4] Y. S. Huang and C. Y. Suen. A Method of Combining Multiple Experts for the Recognition of Unconstrained Handwritten Numerals. *IEEE Trans. on PAMI*, 17(1):90–94, 1995.
- [5] H.-J. Kang. Combining multiple classifiers based on third-order dependency for handwritten numeral recognition. *PRL*, 24(16):3027–3036, 2003.
- [6] H.-J. Kang, K. Kim, and J. H. Kim. Optimal Approximation of Discrete Probability Distribution with *k*th-order Dependency and Its Applications to Combining Multiple Classifiers. *PRL*, 18(6):515–523, 1997.
- [7] H.-J. Kang and S.-W. Lee. Combining Classifiers based on Minimization of a Bayes Error Rate. In *Proc. of the 5th IC-DAR*, pages 398–401, 1999.
- [8] T. Matsui, T. Tsutsumida, and S. N. Srihari. Combination of Stroke/Background Structure and Contour-direction Features in Handprinted Alphanumeric Recognition. In *Proc. of the 4th IWFHR*, pages 87–96, 1994.
- [9] I.-S. Oh and C. Y. Suen. Distance features for neural network-based recognition of handwritten characters. *IJ-DAR*, 1(2):73–88, 1998.
- [10] C. Y. Suen, C. Nadal, R. Legault, T. A. Mai, and L. Lam. Computer Recognition of Unconstrained Handwritten Numerals. *Proc. of IEEE*, pages 1162–1180, 1992.
- [11] D. C. C. Wang and A. K. C. Wong. Classification of Discrete Data with Feature Space Transform. *IEEE TAC*, AC-24(3):434–437, 1979.
- [12] L. Xu, A. Krzyzak, and C. Y. Suen. Methods of Combining Multiple Classifiers and Their Applications to Handwriting Recognition. *IEEE TSMC*, 22(3):418–435, 1992.