Classifier Combining Rules Under Independence Assumptions*

Shoushan Li and Chengqing Zong

National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100080, China
{sshanli,cqzong}@nlpr.ia.ac.cn

Abstract. Classifier combining rules are designed for the fusion of the results from the component classifiers in a multiple classifier system. In this paper, we firstly propose a theoretical explanation of one important classifier combining rule, the sum rule, adopting the Bayes viewpoint under some independence assumptions. Our explanation is more general than what did in the existed previous by Kittler *et al.* [1]. Then, we present a new combining rule, named SumPro rule, which combines the sum rule with the product rule in a weighted average way. The weights for combining the two rules are tuned according to the development data using a genetic algorithm. The experimental evaluation and comparison among some combining rules are reported, which are done on a biometric authentication set. The results show that the SumPro rule takes a distinct advantage over both the sum rule and the product rule. Moreover, this new rule gradually outperforms the other popular trained combining rules when the classifier number increases.

Keywords: Pattern Classification, Multiple Classifier System, Combining Rules.

1 Introduction

Combining multiple classifiers is a learning method where a collection of a finite number of classifiers is trained for the same classification task. Over the past years, this method has been considered as a more practical and effective solution for many recognition problems than using one individual classifier [1] [2].

An important issue in combining classifiers is that of the combining rules which are designed to fuse the results from the component classifiers. Generally, the combining rules are usually categorized into two categories: fixed rules and trained rules. The fixed rules combine the classification results in some fixed mode independent of the application tasks, notably the sum rule and the product rule [1]. And the trained rules combine the results in a trained way, such as weighted sum rule [3], Behavior-Knowledge Space algorithm [4], Decision Template method [5] and Dempster-Shafer (DS) [2] [6] method. Some related experimental studies contribute

* This work has been partially funded by the Natural Science Foundation of China under Grant Nos. 60575043 and 60121302.

M. Haindl, J. Kittler, and F. Roli (Eds.): MCS 2007, LNCS 4472, pp. 322–332, 2007.

to the comparison between these two kinds of methods [7] [8]. Given their extensive experimental results, it is still difficult to draw a consistent conclusion about which kind or which particular rule performs better than the others. The difficulty mainly lies on the lack of explicit theoretical analysis of these rules. Therefore, some theoretical studies on these rules appear with the objective to explain why some combination methods work better and in what cases they perform better than the others. One important work of Kittler *et al.* [1] develops a common theoretical framework based on the different feature sets, where many fixed combining rules such as the product rule, sum rule, min rule, max rule and vote rule are derived. They report that the sum rule outperforms the other rules because of its resilience to estimation errors.

Although it is known that the fixed rules are obtained under strong assumptions, these assumptions still remain unclear. Furthermore, the sum rule takes favorable position in many experimental results, while the assumption for getting this rule is reported much stronger than the product rule [1]. In this paper, we focus on the combining rules based on the different feature sets. Our objective is to give a new explanation to the sum rule. Moreover, we present a new hybrid rule called SumPro which combine the sum rule with the product rule in a weighted average way.

The remainder of this paper is organized as follows. Section 2 presents our theoretical framework on the combining rules through the Bayes theorem under independence assumptions. In particular, we give the detailed analysis on the sum rule and demonstrate that our explanation for the sum rule is more general than that in Kittler *et al.* [1] (fully described in the **Appendix A**). Then, we propose a new combining rule, named as SumPro rule. In Section 3, the experimental study on one data set is given to compare some combining rules for evaluating the SumPro rule. The conclusions are drawn in Section 4.

2 Our Theoretical Framework

In statistical pattern recognition, a given pattern x is assigned to the i th class w_i among all the m classes $W = \{w_1, ..., w_m\}$ with the maximum posterior probability. In a multiple classifier system, when the pattern x is represented by multiple feature sets, i.e., $x = (x_1, x_2, ..., x_R)$, the pattern belonging to class w_i should satisfy the following equation:

$$i = \arg\max_k \; p(w_k \mid x_1, x_2, ..., x_R) \tag{1}$$

2.1 Product Rule

According to the Bayes theorem, the posteriori probability can be rewritten as

$$p(w_k \mid x_1, ..., x_R) = \frac{p(x_1, ..., x_R \mid w_k) P(w_k)}{p(x_1, ..., x_R)} \tag{2}$$

where, $p(x_1, ..., x_R)$ is the joint probability density.

Assume that the feature sets are statistically independent given the class w_k , i.e.,

$$p(x_1,...,x_R \mid w_k) = \prod_{l=1}^{R} p(x_l \mid w_k) \tag{3}$$

Then the posteriori probability can be rewritten as

$$p(w_k \mid x_1,...,x_R) = \frac{p(w_k)\prod_{l=1}^{R} p(x_l \mid w_k)}{p(x_1,...,x_R)} \tag{4}$$

In terms of the posteriori probabilities yielded by the individual classifiers, we obtain the decision rule

$$p(w_k \mid x_1,...,x_R) = p^{-(R-1)}(w_k)\prod_{l=1}^{R} p(w_k \mid x_l)\cdot\frac{\prod_{l=1}^{R} p(x_l)}{p(x_1,...,x_R)} \tag{5}$$

Excluding the same factor of $\frac{\prod_{l=1}^{R} p(x_l)}{p(x_1,...,x_R)}$ for all the classes, we obtain the product rule

$$i = \arg\max_k \; p^{-(R-1)}(w_k)\prod_{l=1}^{R} p(w_k \mid x_l) \tag{6}$$

2.2 Sum Rule

In this section, we give the sum rule in two steps. First, we consider the case in which the system consists of two classifiers for combining ($R = 2$).

Note that the posteriori probability $p(w_k \mid x_1, x_2)$ can also be computed by the probability of $p(\overline{w_k} \mid x_1, x_2)$ as follows

$$p(w_k \mid x_1, x_2) = 1 - p(\overline{w_k} \mid x_1, x_2), \quad \overline{w_k} = W - \{w_k\} \tag{7}$$

Assume that the feature sets are statistically independent given the class set $\overline{w_k}$, i.e.,

$$p(x_1,...,x_R \mid \overline{w_k}) = \prod_{l=1}^{R} p(x_l \mid \overline{w_k}) \tag{8}$$

With the analogous operation for getting formula (5), we get

$$p(w_k \mid x_1, x_2) = 1 - p^{-1}(\overline{w_k})\, p(\overline{w_k} \mid x_1)\, p(\overline{w_k} \mid x_2)\cdot\lambda_2 \tag{9}$$

Where $\lambda_2 = \frac{p(x_1)p(x_2)}{p(x_1, x_2)}$.

Since the sum of $p(\overline{w_k} \mid x_j)$ and $p(w_k \mid x_j)$ equals one, formula (9) becomes

$$p(w_k \mid x_1, x_2)\cdot[1 - p(w_k)] = [1 - p(w_k)] - \lambda_2[1 - p(w_k \mid x_1)][1 - p(w_k \mid x_2)] \tag{10}$$

Applying formula (5), we expand the left of the above formula

$$\begin{aligned} &p(w_k \mid x_1, x_2) - \lambda_2 p(w_k \mid x_1)\, p(w_k \mid x_2) \\ &= 1 - p(w_k) - \lambda_2[1 - p(w_k \mid x_1)][1 - p(w_k \mid x_2)] \end{aligned} \tag{11}$$

With the further simplification, we get

$$p(w_k \mid x_1, x_2) = [1 - p(w_k) - \lambda_2] + \lambda_2 [p(w_k \mid x_1) + p(w_k \mid x_2)] \quad (12)$$

The above formula shows that the combining posterior probability can be expressed as the sum of the individual probabilities in the two-classifier case. And the assumptions (3) and (8) are used in the deducing process.

Secondly, let us consider the case that the system consists of more than two classifiers ($R > 2$).

We can regard the ensemble of former $R-1$ component classifiers as one classifier. Then, according to formula (12), we get

$$p(w_k \mid x_1, ..., x_R) = [1 - p(w_k) - \lambda_R] + \lambda_R [p(w_k \mid x_1, ..., x_{R-1}) + p(w_k \mid x_R)] \quad (13)$$

Where, $\lambda_R = \dfrac{p(x_1, ..., x_{R-1}) p(x_R)}{p(x_1, ..., x_R)}$.

By expanding the above formula, we get the following expression

$$p(w_k \mid x_1, ..., x_R) = \sum_{l=2}^{R} [(\prod_{q=l}^{q=R} \lambda_q) \cdot p(w_k \mid x_l)] + \prod_{q=2}^{q=R} \lambda_q \cdot p(w_k \mid x_1) + cont. \quad (14)$$

Where, $\lambda_q = \dfrac{p(x_1, ..., x_{q-1}) p(x_q)}{p(x_1, ..., x_q)}$, and $cont.$ is the remaining terms which is only related to the class prior probability $p(w_k)$.

Formula (14) demonstrates that the classifiers using independent feature sets should be combined in a linear weighted way under the two assumptions (3) and (8). Since our objective is to get the sum rule, we would cut off the weights.

We assume that the feature measurements x_j $(j = 1, 2, ..., R)$ are statistically independent, then the value of λ_j $(j = 1, 2, ..., R)$ equals one, i.e.,

$$\lambda_j = 1 \ (j = 1, 2, ..., R) \quad (15)$$

Under this assumption, the sum rule for multiple classifiers can be conclude from formula (14) that

$$i = \arg\max_k \{-(m-1) p(w_k) + \sum_{l=1}^{R} p(w_k \mid x_l)\} \quad (16)$$

This formula implies that the prediction decision can be drawn according to the sum of the posterior probabilities yielded by the individual classifiers. This is the same sum rule that has been widely used in the multiple classifier system field. Compared to the product rule, two more independent assumptions, (8) and (15), are involved in the sum rule. To further understand these assumptions, some comments about these two rules are described as follows.

- The assumptions (3) and (8) are two conditional dependent assumptions and they are popularly used in pattern recognition literature (e.g., Naïve Bayes classifier) for simplifying the analysis. The difference between these two assumptions lies in their different conditions, i.e., one has the given class w_k while the other has the given class set $\overline{w_k}$. Note that, in two-class case, when the class set $\overline{w_k}$

merely consists of one class, these two assumptions are equivalent. It is also interesting to point out that this special case is a good explanation to one conclusion drawn in the previous work of Tax *et al.* [9], which states that, in a two-class problem, the sum rule and the product rule achieve comparable performance [9].

- Another assumption (15) involved in the sum rule is so strong that it would be violated in many applications. However, this independent assumption is actually a sufficient condition for the sum rule in our explanation. As shown in (12), in two-classifier case, this assumption is needless for the sum rule under equal prior assumption.
- Another important related work have been given by Kittler *et al.* [1], who also deduced the sum rule with a rather strong assumption. This assumption states that the posteriori probabilities computed by the respective classifiers will not deviate dramatically from the prior probabilities. It is not difficult to prove that our explanation is more general than theirs. A detailed proof can be found in the **Appendix A**.
- We must concede that to satisfy all the assumptions at the same time is really difficult in many applications. Nevertheless, the sum rule performs so well in error sensitivity that it outperforms other fixed rules in many experimental results [1] [10].
- Other common fixed rules, such as the max rule, the min rule and the vote rule can be easily obtained using the above two basic rules [1]. These rules are discussed in detail in Kittler *et al.* [1].

2.3 SumPro Rule

As mentioned above, the sum rule requires much stronger assumptions than the product rule but it takes lower error sensitivity than the product rule. In real problems (when approximated posteriors are used), it is interesting to look for a hybrid method of these two rules which could combine the strengths of the product and sum rules. In this section, we propose a new combining rule called the SumPro rule.

Firstly, let's consider the two-classifier case ($R = 2$).

Under the assumption (3) and (15), we get

$$p(w_k \mid x_1, x_2) = p^{-1}(w_k)\, p(w_k \mid x_1)\, p(w_k \mid x_2) \tag{17}$$

Under the assumption (15), formula (12) can be simplified as

$$p(w_k \mid x_1, x_2) = -p(w_k) + p(w_k \mid x_1) + p(w_k \mid x_2) \tag{18}$$

Thus, we have

$$\begin{aligned} p(w_k \mid x_1, x_2) = (1-\omega)\cdot[-p(w_k) + p(w_k \mid x_1) + p(w_k \mid x_2)] \\ + \omega \cdot p^{-1}(w_k)\, p(w_k \mid x_1)\, p(w_k \mid x_2) \end{aligned} \tag{19}$$

Where ω $(0 \le \omega \le 1)$, can be regard as a variable varying from zero to one.

Note that the sum rule and product rule expressions become the special cases when $\omega = 0$ and $\omega = 1$ respectively. In real applications, it is always possible to find a suitable value of ω under some criterion.

As to multiple-classifier case, the posterior probability can be computed by using following formula iteratively

$$\begin{aligned} p(w_k \mid x_1,...,x_R) = (1-\omega_{R-1})\cdot[-p(w_k) + p(w_k \mid x_1,...,x_{R-1}) + p(w_k \mid x_R)] \\ + \omega_{R-1}\cdot p^{-1}(w_k)\, p(w_k \mid x_1,...,x_{R-1})\, p(w_k \mid x_R) \end{aligned} \tag{20}$$

Then, the SumPro rule can be defined as following

$$\begin{aligned} i = \arg\max_k \{(1-\omega_{R-1})\cdot[-p(w_k) + p(w_k \mid x_1,...,x_{R-1}) + p(w_k \mid x_R)] \\ + \omega_{R-1}\cdot p^{-1}(w_k)\, p(w_k \mid x_1,...,x_{R-1})\, p(w_k \mid x_R)\} \end{aligned} \tag{21}$$

In real applications, one essential task is to find a way for training the values of ω_i $(i = 1,...,R-1)$. One simple way is an exhaustively search for the optimal values on the training set under some criterion. However, this is impracticable when the classifier's number is large. Considering that the genetic algorithm is a good tool for optimization problems [11], we apply genetic algorithm to tune the values of ω_i $(i = 1,...,R-1)$ according to some optimization criterion.

3 Empirical Study

In this section, we perform experimental study to compare the combining rules on a biometric authentication data set. A biometric authentication system is designed to verify the identity of a person based on biometric measures such as the person's face, voice, iris or fingerprints [10]. Use of multiple biometric indicators, known as multimodal biometrics, has been shown to increase the authentication accuracy [12]. A number of combining rules have been applied to combine the results of the multiple biometric indicators, where the sum rule, the DS rule, and the support vector machine (SVM) rule are usually reported as the champions [1] [13] [14].

Our experimental data set[1] is presented by Poh and Bengio [15] to encourage researchers to focus on the problem of biometric authentication score-level fusion. The scores are taken from the XM2VTS database, which contains video and speech data from 295 subjects. There exist two configurations called Lausanne Protocol I (LP1) and Protocol II (LP2) in the dataset with different partitioning approaches of the training and development sets. In both configurations, the test set is the same [15]. In our experiment, we pick the LP1 for our experimental study where eight different classifiers are available.

Two kinds of errors occur in a biometric authentication system, i.e., false acceptance of the impostors and false rejection of the clients. Correspondingly, there are two measures commonly used for evaluating the system, i.e., false acceptance rate (FAR) and false rejection rate (FRR). Generally, FAR and FRR are balanced by the threshold which is used for determining whether one person is an impostor or a client. Another evaluation criteria is defined as the mean value of FAR and FRR, called Half Total Error Rate ($HTER$), i.e., $HTER = (FAR + FRR)/2$ [15].

[1] It is available at http://www.idiap.ch/~norman/fusion/main.php?bodyfile=entry_page.html.

Table 1. Comparison of performances using different combining rules

Rules	N=2	N=3	N=4	N=5	N=6	N=7	N=8
Product	2.36	1.86	1.56	1.43	1.74	2.54	3.50
Sum	2.09	1.39	1.08	0.91	0.83	0.78	0.75
Max	2.51	2.02	1.76	1.52	1.36	1.23	1.13
Min	3.04	2.83	2.74	2.64	2.51	2.41	2.35
SumPro	2.03	1.28	0.82	0.74	**0.62**	**0.58**	**0.31**
wighted sum	1.82	1.15	0.88	0.76	0.70	0.66	0.63
DS	1.47	1.20	1.07	0.99	0.90	0.82	0.76
SVM	**1.44**	**0.90**	**0.74**	**0.68**	0.64	0.60	0.52

The development set is used to estimate both the threshold value for rejecting and the approximate optimal omega values ω_i $(i = 1,..., R-1)$ for the SumPro rule by the genetic algorithm. In our experiment, the genetic operators, including selection, crossover, and mutation are all set to the default values in the GA tool in Matlab 7.0. And the optimization criterion is to obtain the best $HTER$ value in the development set.

The best and the worst $HTER$ values over all the single classifiers are 1.53% and 7.60%. We perform the combining methods by combining all the possible combinations of N $(N = 2,3,...,8)$ classifiers taken from the eight ones. The mean $HTER$ values of every aggregate of the N classifiers are shown in **Table 1**. Below we highlight some of our interesting findings from **Table 1**.

First, we find that combining classifier can contribute to the overall performance of the authentication system since, in the eight-classifier case, results with most combining rules are better than the best result of the single classifiers. In general, good trained rules usually outperform the fixed rules. In our experiment, the weighted sum rule, the DS rule and the SVM rule obtain superior results to the fixed rules, such as the product rule, the max rule and the min rule. However, the sum rule, as a fixed rule, achieves comparable performance with the trained rules.

Then, the SumPro rule is consistently preferable than both the sum rule and the product rule. This conclusion is encouraging because the sum rule has been reported as the best rule in many related studies. Moreover, this new rule gradually outperforms the other popular trained combining rules when the classifier number increases (when $N = 6,7,8$).

4 Conclusion

In summary, the contribution of this paper is twofold. At first, we give a theoretical analysis on the two fixed rules of the product and the sum rule which are often seen as two basic rules in the fixed rules. The proposed assumptions can help us to better understand the fixed rules. Second, we present a new combining rule called SumPro rule which combines the product rule with the sum rule. Experimental results reveal this new rule performs particularly well when the number of the classifiers is large.

References

1. J. Kittler, M. Hatef, R.P.W. Duin, and J. Matas, On Combining Classifiers. PAMI, vol.20, pp.226-239, 1998
2. L. Xu, A. Krzyzak, and C.Y. Suen, Methods of Combining Multiple Classifiers and Their Applications to Handwriting Recognition. IEEE Tran. Systems, Man and Cybernetics, vol.22(3), pp.418-435, May/June. 1992
3. G. Fumera, and F. Roli, A Theoretical and Experimental Analysis of Linear Combiners for Multiple Classifier Systems. IEEE Trans. PAMI, vol.27, pp.942 – 956, 2005
4. Y.S. Huang, and C.Y. Suen, A Method of Combining Multiple Experts for the Recognition of Unconstrained Handwritten Numerals. IEEE Trans. PAMI, vol.17(1), pp.90-94, 1995
5. L. I. Kuncheva, J. C. Bezdek, R. P.W. Duin, Decision templates for multiple classifier fusion: an experimental comparison. Pattern Recognition, vol.34, pp.299-314, 2001
6. Y. Sugie, and T. Kobayashi, Media-integrated Biometric Person Recognition Based on the Dempster-Shafer Theory. 16th International Conference on Pattern Recognition (ICPR), vol.4, pp.381-384, 2002
7. R.P.W. Duin, and D.M.J. Tax, Experiments with Classifier Combining rules. Proceedings of First International Workshop on Multiple Classifier System (MCS 2000), Lecture Notes in Computer Science, vol.1857 pp.16-29, 2000
8. F. Roli, S. Raudys, and G.L. Marcialis, An Experimental Comparison of Fixed and Trained Fusion Rules for Crisp Classifiers Outputs. Proceedings of First International Workshop on Multiple Classifier System (MCS 2002), Lecture Notes in Computer Science, vol.2364, 2002
9. D.M.J. Tax, M. van Breukelen, R.P.W. Duin, and J. Kittler, Combining Multiple Classifiers by Averaging or by Multiplying. Pattern Recognition, vol.33, pp.1475-1485, 2000
10. A. Ross and A.K. Jain, Information Fusion in Biometrics. Pattern Recognition Letters, vol.24, no.13, pp.2115-2125, 2003
11. D. E. Goldberg, Genetic Algorithm in Search, Optimization and Machine Learning, Addison Wesley, Reading, 1989
12. R. Snelick, U. Uludag, A. Mink, M. Indovina, and A. Jain, Large-Scale Evaluation of Multimodal Biometric Authentication Using State-of-the-Art Systems. IEEE Trans. PAMI, vol.27, no.3, pp.450-455, Mar., 2005
13. S. Ben-Yacoub, Y. Abdeljaoued, and E. Mayoran, Fusion of Face and Speech Data for Person Identity Verification. IEEE Trans. on Neural Networks, vol.10(5), pp.1065-1074, 1999
14. K. Chang, and K. W. Bowyer, Comparison and Combination of Ear and Face Images in Appearance-Based Biometrics. IEEE Trans. PAMI, vol.25, no.9, pp.1160-1165, 2003
15. N. Poh, and S. Bengio, A Score-level Fusion Benchmark Database for Biometric Authentication. AVBPA, Lecture Notes in Computer Science, vol.3546, pp.1059-1070, 2005

Appendix A

In this appendix, we demonstrate that our explanation for the sum rule is more general than the explanation in Kittler *et al.* [1]. In Kittler *et al.* [1], they presented an explanation of sum rule under the assumption that the posteriori probabilities

computed by the respective classifiers will not deviate dramatically from the prior probabilities, i.e.,

$$p(w_k \mid x_j) = p(w_k)(1+\delta_{kj}), \tag{22}$$

where δ_{kj} satisfies $\delta_{kj} << 1$. First, let us prove the following theorem.

Theorem 1. *Under the assumption (3), the assumption (22) is a sufficient condition for the assumption (8) and the assumption (15).*

Proof. ***(a)*** Firstly, we consider the assumption (15).

Substituting (22) into (3) and applying the Bayes theory, we get

$$p(x_1,...,x_r \mid w_j) = \prod_{l=1}^{r} p(x_l) + \sum_{l=1}^{r} \delta_{jl} \cdot \prod_{l=1}^{r} p(x_l) \tag{23}$$

Because of the basic character of the probability, we have

$$\sum_{j=1}^{m} p(w_j \mid x_l) = 1 \tag{24}$$

Thus,

$$\sum_{j=1}^{m} p(w_j)(1+\delta_{jl}) = 1 \tag{25}$$

Because $\sum_{j=1}^{m} p(w_j) = 1$, we can find

$$\sum_{j=1}^{m} [p(w_j) \cdot \delta_{jl}] = 0 \tag{26}$$

According to the law of the total probability,

$$p(x_1,...,x_r) = \sum_{j=1}^{m} p(w_j)\, p(x_1,...,x_r \mid w_j) \tag{27}$$

Substituting (23) into (27), we find

$$p(x_1,...,x_r) = \prod_{l=1}^{r} p(x_l) + \sum_{j=1}^{m} \{p(w_j) \cdot [\sum_{l=1}^{r} \delta_{jl} \cdot \prod_{l=1}^{r} p(x_l)]\} \tag{28}$$

From formula (26), we have

$$\sum_{j=1}^{m} \{p(w_j) \cdot [\sum_{l=1}^{r} \delta_{jl} \cdot \prod_{l=1}^{r} p(x_l)]\} = 0 \tag{29}$$

Substituting (29) into (28), we obtain

$$p(x_1,...,x_r) = \prod_{l=1}^{r} p(x_l) \tag{30}$$

Formula (30) exactly implies that each feature set x_j $(j = 1,2,...,R)$ is independent from each other, thus the assumption (15) can be satisfied.

(b) Then, we consider the assumption (8).

As discussed in Section 2.1, from formula (5) and the assumption (15), we get

$$p(w_k \mid x_1, x_2, ..., x_r) = p^{-1}(w_k)\prod_{j=1}^{r} p(w_k \mid x_j) \tag{31}$$

Substituting (22) into the above formula, we obtain

$$p(w_k \mid x_1, x_2, ..., x_n) = p(w_k)(1+\sum_{j=1}^{R} \delta_{kj}) \tag{32}$$

Thus,

$$p(\overline{w_k} \mid x_1, x_2, ...x_n) = 1 - p(w_k \mid x_1, x_2, ..., x_n) = 1 - p(w_k)(1+\sum_{j=1}^{R} \delta_{kj}) \tag{33}$$

On the other hand, when there are two classifiers for combination

$$\begin{aligned} p^{-1}(\overline{w_k})p(\overline{w_k} \mid x_1)p(\overline{w_k} \mid x_2) &= p^{-1}(\overline{w_k})[(1-p(w_k \mid x_1)][1-p(w_k \mid x_2)] \\ &= p^{-1}(\overline{w_k})[(1-p(w_k))(1-p(w_k))(1+\delta_{k1}+\delta_{k2})] \\ &= 1-p(w_k)(1+\delta_{k1}+\delta_{k2}) \end{aligned} \tag{34}$$

As to multiple classifiers, it is not difficult to get

$$p^{-(R-1)}(\overline{w_k})\prod_{l=1}^{R} p(\overline{w_k} \mid x_l) = 1 - p(w_k)(1+\sum_{j=1}^{R} \delta_{kj}) \tag{35}$$

From (32) and (35), we find

$$p(\overline{w_k} \mid x_1, x_2, ...x_n) = p^{-(R-1)}(\overline{w_k})\prod_{l=1}^{R} p(\overline{w_k} \mid x_l) \tag{36}$$

According to the Bayes theorem, the left item and the right item of (36) can be expanded as

$$p(\overline{w_k} \mid x_1, x_2, ...x_n) = \frac{p(x_1, x_2, ..., x_n \mid \overline{w_k})p(\overline{w_k})}{p(x_1, x_2, ..., x_n)} \tag{37}$$

and,

$$p^{-(R-1)}(\overline{w_k})\prod_{l=1}^{R} p(\overline{w_k} \mid x_l) = \frac{p(\overline{w_k})\prod_{l=1}^{R} p(x_l \mid \overline{w_k})}{\prod_{l=1}^{R} p(x_l)} \tag{38}$$

From (36), (37) and (38), we get

$$\frac{p(x_1, x_2, ..., x_n \mid \overline{w_k})p(\overline{w_k})}{p(x_1, x_2, ..., x_n)} = \frac{p(\overline{w_k})\prod_{l=1}^{R} p(x_l \mid \overline{w_k})}{\prod_{l=1}^{R} p(x_l)} \tag{39}$$

Applying (30) to the above formula, we get

$$p(x_1, x_2, ..., x_n \mid \overline{w_k}) = \prod_{l=1}^{R} p(x_l \mid \overline{w_k}) \tag{40}$$

The above expression is exactly the assumption (8).

Therefore, we can conclude that the assumption (22) is a sufficient condition for the assumption (8) and (15).

In our explanation, under the equal prior assumption, the sum rule can obtained by only using the assumption (3) when there exist two class labels and two classifiers (see formula (12) and the assumption (8) is the same as the assumption (3) in two-class problem). In other words, the assumption (22) is unnecessary for getting the sum rule in this special case.

Considering the **Theorem 1** and the special case above, we can conclude that our explanation with the independence assumptions is more general than the explanation by Kittler *et al.* [1].