Improved computation of beliefs based on confusion matrix for combining multiple classifiers

L. Chen and H.L. Tang

One approach among a number of strategies for multiple classifier combination is to calculate the beliefs of confidence based on confusion matrices from individual classifiers. To achieve precise belief computation, based on previous researchers' work, presented is a better algorithm with a more generic capability, showing improved performance.

Introduction: Applications in the area of pattern recognition are increasing combining multiple classifiers instead of developing the 'best' classifier [1, 4, 5]. Xu *et al.* [1] summarise three types of combination problems according to three levels of output information by various classifiers: the abstract level (a classifier only outputs a unique label), the rank level (ranked labels are output by a classifier) and the measurement level (the posterior probability is known for each classifier). A majority-voting algorithm is the main method applied in the abstract level. This is normally applied in the case that a single classifier only outputs one label and posterior probability is unknown. An alternative approach [1, 2] is to calculate the beliefs based on confusion matrices of different classifiers and these beliefs will provide the basic information for multiple classifier combination. The described method for computing beliefs [1, 2] does not, however, consider the effect of the numbers of test data in individual classes in a confusion matrix, and it is assumed that in each class there is an equal number of test data. A similar problem exists [3] where the different numbers of test data in individual classes lead to the wrong beliefs in measuring the similarity between classes and further imprecise semantic reasoning that is based on such beliefs.

In this Letter we present a theoretical method to compute beliefs with a better degree of precision, which produces better experimental results in practice compared with the published algorithms [1–3]. The remaining part of this Letter is arranged as follows: the computation methods of beliefs based on the confusion matrix of a single classifier are reviewed, followed by a proposal for an improved algorithm. The sum combination rule is applied to compare the previous and the improved algorithms. Experimental results are discussed and conclusion is given.

Confusion matrix, beliefs and combination: When posterior probability is not available for combining multiple classifiers, a natural solution is to vote for the final decision by considering the labels directly outputted from all the classifiers. Such a method is, however, based on such an assumption: most or all classifiers achieve good accuracy, so this ignores the detailed performance information of every single classifier. An alternative is to calculate beliefs based on confusion matrices of individual classifiers and then combine them based on these beliefs. A confusion matrix is a detailed report on the performance of a single classifier. Let us assume that M classes in pattern space Z with K classifiers. A classifier is a black box or a function:

$$e_k(x) = j, \quad k = 1, 2, \ldots, K, \quad j \in \{1, 2, \ldots, M, M+1\}$$

and its confusion matrix is:

$$CM_k = \begin{pmatrix} n_{11}^k & n_{12}^k & \ldots & n_{1M}^k & n_{1(M+1)}^k \\ n_{21}^k & n_{22}^k & \ldots & n_{2M}^k & n_{2(M+1)}^k \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ n_{M1}^k & n_{M2}^k & \ldots & n_{MM}^k & n_{M(M+1)}^k \end{pmatrix},$$

$M+1$ is an unknown label

Each row i corresponds to class i and each column j corresponds to $e_k(x) = j$. The element n_{ij}^k means that n_{ij}^k samples of class i are assigned to class j by $e_k(x)$. CM_k is obtained by executing $e_k(x)$ on the test dataset after $e_k(x)$ is trained. The number of samples in class i is: $n_{i.}^k = \sum_{j=1}^{M+1} n_{ij}^k$, where $i = 1, 2, \ldots, M$, and the number of samples labelled j by $e_k(x)$ is $n_{.j}^k = \sum_{i=1}^{M} n_{ij}^k$, where $j = 1, 2, \ldots, M+1$.

Table 1: Experimental results of individual classifiers and averaging methods by applying formulas (1) and (2) on 63 classes (for demonstration purpose, randomly chosen classes are displayed in this Letter)

Class	#1 Accuracy (%)	#2 Accuracy (%)	#3 Accuracy (%)	Average (1) Accuracy (%)	Average (2) Accuracy (%)
1. Adipose tissue	51.85	**59.26**	54.31	57.50	**59.26**
...	...	...	...	...	...
5. Anus epithelium	53.33	40.00	36.67	53.33	**66.67**
6. Anus lamina propria	**97.06**	94.12	88.24	94.12	94.12
7. Some appendix glands next to lumen	56.76	54.05	43.24	51.35	**70.27**
8. Serosa next to muscularis externa	**60.00**	40.00	**60.00**	**60.00**	**60.00**
9. Muscularis mucosae	83.33	66.67	66.67	66.67	**100**
...	...	...	...	...	...
22. Junction between submucosa and tight muscle	57.14	42.86	50.00	42.86	**64.29**
23. Lymph nodule	85.19	79.63	85.19	83.33	**90.74**
24. Lamina propria	60.00	55.00	**65.00**	55.00	**65.00**
...	...	...	...	...	...
47. Small intestine: junction:intestinal glands and lamina propria	20.00	0	0	0	**40.00**
48. Small intestine: junction:intestinal glands and muscularis mucosae	51.06	57.45	46.81	57.45	**61.70**
49. Small intestine: junction: lumen and villi	28.21	41.03	38.46	41.03	**48.72**
...	...	...	...	...	...
57. Stomach foveolae(long)	45.46	50.00	22.73	45.46	**63.64**
58. Stomach foveolae(middle)	58.93	50.00	51.79	50.00	**67.86**
59. Stomach foveolae(surface)	59.21	**65.79**	64.47	**65.79**	61.84
60. Stomach: junction:fundus glands and lamina propria	72.73	63.64	61.11	63.64	**81.82**
61. Stomach: junction:fundus glands and muscularis mucosae	50.00	40.00	50.00	30.00	**60.00**
62. Stomach: junction:lumen and focal oedema	62.50	50.00	**75.00**	50.00	50.00
63. Stomach: junction:lumen and foveolae	61.91	**69.05**	64.29	**69.05**	**69.05**
Total accuracy	70.73	69.66	69.74	70.37	**72.27**

In [1], the belief $b_k(\bullet)$ is calculated as follows:

$$b_k(x \in class\ i | e_k(x) = j, EN) = P(x \in class\ i | e_k(x) = j, EN) = \frac{n_{ij}^k}{n_{.j}^k}, \quad i, j = 1, 2, \ldots, M+1 \quad (1)$$

EN denotes the common classification environment. However this computation is only suitable in cases when the number of samples in each class is the same, i.e. $n_{i.}^k = n_{j.}^k$, $(i \neq j)$. When $n_{i.}^k \gg n_{j.}^k$ $(i \neq j)$, even if the accuracy for the classifier e_k $(x \in class\ i)$ is high, the small number of misclassification, i.e. e_k $(x \in class\ i) = j$, where $i, j \in \{1, 2, \ldots, M+1\}$ and $i \neq j$, will cause the imprecise beliefs. In the proposed method, (1) is modified into:

$$b_k(x \in class\ i | e_k(x) = j, EN) = P(x \in class\ i | e_k(x) = j, EN) = \frac{n_{ij}^k / n_{i.}^k}{\sum_{t=1}^{M} (n_{tj}^k / n_{t.}^k)} = \frac{n_{ij}^k}{\sum_{t=1}^{M} (n_{tj}^k \varsigma_{it}^k)}, \quad (2)$$

$$where\ \varsigma_{it}^k = \frac{n_{i.}^k}{n_{t.}^k}, \quad i, j = 1, 2, \ldots, M+1$$

When the difference between the numbers of test data for different classes becomes large, (1) is less precise than (2) in computing the

beliefs for individual classifiers. The imprecise beliefs will give the wrong measurement for multiple classifier combination process.

When multiple classifiers $e_1, e_2, \ldots, e_K$ are developed, their correspondent beliefs $b_1, b_2, \ldots, b_K$ are computed based on the performance of base classifiers. It is assumed that individual classifiers are mutually exclusive. Typical combination strategies are averaging and multiplying algorithms. A theoretical and practical comparison has been discussed [5] for these two types of methods, the conclusion being that averaging-estimated posterior probabilities are to be preferred when posterior problems are not well estimated. Pathological images [6] have been used to test the proposed approach. The posterior probabilities from confusion matrices of individual classifiers are not well estimated (see the performance of base classifiers in Table 1) due to the limitations of current image processing techniques and complicated nature of pathological images, so the following average algorithm (3) has been employed:

$$\begin{aligned} b(i) &= b(x \in class\ i | e_1(x), e_2(x), \ldots, e_K(x), EN) \\ &= P(x \in class\ i | e_1(x), e_2(x), \ldots, e_K(x), EN) \\ &= \frac{1}{M+1} \sum_{k=1}^{K} b_k(x \in class\ i | e_k(x), EN), \quad i = 1, 2, \ldots, M+1 \end{aligned} \tag{3}$$

The belief of making the final decision to assign x to *class* j ($j = 1, 2, \ldots, M+1$) is $B(j) = \max_{i=1}^{M+1} b(i)$.

Experiment results: In this work, the pathological image collection and its classes (63 classes) [6] are used as the dataset, which is randomly divided into three subsets without considering the numbers of the test data in individual classes: training dataset (2754 samples), test dataset1 (2755 samples), testing dataset2 (2528 samples). To avoid inaccurate belief estimations due to over-training, the training dataset is used to train individual classifiers, none-overlapping test dataset1 is used to produce the confusion matrices of base classifiers, and the second test dataset is applied to test the aforementioned idea. The classification algorithm is based on multi-class support vector machines (SVMs). Three multi-class SVMs are trained based on colour histogram (#1), texture feature extracted by Gabor filters (#2), texture feature by wavelet (#3). The experimental results are shown in Table 1. The first three columns (#1–#3) demonstrate the performance of the individual classifiers. The accuracy of the base classifiers is not high enough to be reliable, implying that the posterior probability is not well estimated; therefore the averaging combination strategy is adopted as discussed above. The formulas (1) and (2) are applied into the average algorithm (3), respectively, and the detailed comparison is reported in the last two columns of the Table 1. The averaging method using the beliefs calculated in (1) results in even worse performance than some of the base classifiers; while the averaging combination strategy based on the improved algorithm (2) achieves the best performance in categorising most of the classes as well as in the total accuracy.

Conclusion: The aim of this work was to produce a general and precise computation method of beliefs based on confusion matrices of individual classifiers, which will serve as the basic information for multiple classifier combination at the abstract level [1, 2] as well as providing knowledge for further semantic reasoning [3]. We improved the previous algorithms [1–3] and used a large-scale pathological image database to test the proposed idea with the improved results.

Acknowledgment: L. Chen is part-supported by the Overseas Research Students Awards Scheme, UK.

5 November 2003
Electronics Letters online no: 20040176
doi: 10.1049/el:20040176

L. Chen and H.L. Tang (*Department of Computing, University of Surrey, Guildford, Surrey GU2 7XH, United Kingdom*)

References

1 Xu, L., Kryzak, A., and Suen, C.V.: 'Methods of combining multiple classifiers and their applications to handwriting recognition', *IEEE Trans. Syst. Man Cybern.*, 1992, **22**, (3), pp. 418–435
2 Parker, J.R.: 'Rank and response combination from confusion matrix data', *Inf. Fusion*, 2001, **2**, pp. 113–120
3 Tang, H.L.: 'Semantic analysis of image content for intelligent retrieval and automatic annotation of medical images'. PhD thesis, University of Cambridge, 2000
4 Kittler, J., *et al.*: 'On combining classifiers', *IEEE Trans. Pattern Anal. Mach. Intell.*, 1998, **20**, (3), pp. 226–239
5 Tax, D., *et al.*: 'Combining multiple classifiers by averaging or by multiplying?', *Pattern Recognit.*, 2000, **33**, pp. 1475–1485
6 Tang, H.L., Hanka, R., and Ip, H.S.H.: 'Histological image retrieval based on semantic content analysis', *IEEE Trans. Inf. Technol. Biomed.*, 2003, **7**, (1), pp. 26–36