Multiclass Probabilistic Classification for Support Vector Machines

Ji-Sang BAE†, Nonmember and Jong-Ok KIM†a), Member

SUMMARY Support Vector Machine (SVM) is one of the most widely used classifiers to categorize observations. This classifier deterministically selects a class that has the largest score for a classification output. In this letter, we propose a multiclass probabilistic classification method that reflects the degree of confidence. We apply the proposed method to age group classification and verify the performance. key words: multiclass classification, probabilistic classification, SVM, age-group classification

1. Introduction

Classification algorithms treat how to categorize observations into several classes. These methods are used to find the most similar class of the unknown observation. Support Vector Machine (SVM) is one of the most widely used classification techniques. Classical SVM computes the score of each class and deterministically selects a class that has the largest score for a classification output [1].

However, producing a probabilistic classification is often useful in many real-life recognition applications instead of deterministic one. This probabilistic classification is beneficial in particular when the outputs of several weak classifiers are combined for an overall decision. For example, in age group classification that we primarily consider in this letter, state-of-the-art methods frequently extract multiple features from a facial image to improve the classification accuracy [2]–[4]. Even though these methods show better results than a single feature [3], [4], their performance can be further improved by training local features with weak classifiers individually. If the combination of individual weak classifiers considers the degree of confidence for each weak classification, the final decision would be more accurate, especially for score-sensitive cases (the individual feature is located at near hyperplane). For our age estimation, the score-sensitive case often happens due to illumination, expression such as laugh, wearing glasses and so on. This motivates us to develop the probabilistic SVM classification.

Meanwhile, other methods have been proposed to map the deterministic SVM outputs to probability. Platt suggested a parametric probability model that uses the statistical distribution of training samples [5]. This method finds the best-fit sigmoid function to the outputs of SVM on empirical data. The work in [6] exploits the Platt’s method [5] to find the posterior probability for very large classification problem. The work in [7] used Gaussian weighted distance between input and the center of each class for probability estimation. Algorithms used in [5], [7] basically convert the score of each class to probability and consider only binary classification.

In this letter, we propose a multiclass probability estimation method that reflects the degree of confidence. To estimate the probability of the unknown observation we use the distance from hyperplane. By considering sign of distance, the probability of each class is separately estimated by the distribution of its own empirical data. This enables us to reduce the effect of score-sensitive cases into an overall decision. We apply the proposed probabilistic classification method to age-group classification and confirm its superiority.

This letter is organized as follows. In Sect. 2, we review SVM distance briefly and illustrate how to estimate probability. Then in Sect. 3, we present the experimental results in 4 class age-group classification and conclude the letter in Sect. 4.

2. The Proposed Method

In this letter, we propose a multiclass probabilistic classification method using SVM that estimates the probability of the unknown observation to each class, \( P(\text{class} \mid \text{unknown observation}) \). To estimate the probability, we use the distance between the unknown observation and hyperplane that separates two classes. Since this distance represents the closeness of the unknown observation to each class, larger distance means that the unknown observation is more probable to belong to the class. We use the distance distribution of training samples to estimate the probability of the unknown observation. By combining the estimated probabilities from every pairwise class, we finally estimate \( P(\text{class} \mid \text{unknown observation}) \) in multiclass classification.

2.1 SVM Distance

SVM is one of the widely used classifiers in many areas. It finds a hyperplane that is expected to have maximum margin. To classify the unknown observation \( (X) \) in \( N \) class classification, SVM gets score of each class using weight
vector obtained from SVM learning with labeled observation.

\[ g_i(X) = w_i^T X + \omega_i, \quad i = 1, \ldots, N \]  

(1)

where \( g_i(X) \), \( w_i \) and \( \omega_i \) are score, weight vector and bias of class \( C_i \), respectively. In classical multiclass SVM, the estimated class is determined by comparing scores of all classes. The signed distance of the unknown observation to hyperplane is given by (see Fig. 1),

\[ d_{ij} = \frac{g_i(X) - g_j(X)}{||w_i - w_j||} \]  

(2)

Larger distance means higher probability to positive class. So, we exploit the distance to estimate classification probability.

2.2 Probability Estimation in Multiclass Classification

The distance in (2) is used to estimate the probability of each class. Unlike previous probabilistic SVM methods, we specially consider the sign of the distance. For instance, consider a 3 class classification example (Fig. 2). In the example, the input \( X \) is close to \( C_1 \) against \( C_2 \), \( C_1 \) against \( C_3 \) and \( C_3 \) against \( C_2 \) with reference to the hyperplane \( H_{12}, H_{13} \) and \( H_{23} \) respectively. The sign of each distance \( (d_{ij}) \) represents which class the input \( X \) belongs to. The closeness to the class is represented by the length of distance from each hyperplane. So it is natural to consider both sign and length of the distance to estimate the probability for multiclass SVM. We use the Cumulative Density Function (CDF) of training samples for each class pair (Fig. 3 (a)). The distances of labeled training samples are plotted separately depending on their sign. In Fig. 3, red asterisks represent the cumulative distribution of class \( C_i \) and blue circles for class \( C_j \). Then, we find the 3-degree polynomial best-fit functions that minimize the mean square error between each CDF and the best-fit function. \( f_{ij} \) is the best-fit function of class \( C_i \) against class \( C_j \) and \( f_{ji} \) is vice versa. The probability of class \( C_i \) against class \( C_j \) (denoted by \( P_{ji} \)) is estimated using the function, \( P_{ij} = f_{ij}(d_{ij}) \) and similarly the probability of class \( C_j \) against class \( C_i \) is given by \( P_{ji} = f_{ji}(d_{ij}) \).

For multiclass classification extension, the binary classification outputs of all class pairs are combined using “pairwise coupling” strategy. This strategy is generally used for multiclass classification where there exist hyperplanes that separate every pairwise class. For example, in N-class classification, the number of pairwise class is \( \sum_{i=1}^{N-1} i = (N-1)N/2 \). We calculate the distance from hyperplane \( (H_{ij}) \) that separate class \( C_i \) and \( C_j \) for every pairwise class. The probability of class \( C_i \) is estimated by merging the estimated probabilities between class \( C_i \) and all the other classes. This is given by
Given by, the optimal classification is generally expressing the probability for each class. Then, they are merged for final classification. As shown in Fig. 4, different practical usefulness is demonstrated in the experimental probabilistic classifier to the actual classification problem, and an overall decision. Class features are classified individually and combined for probabilistic classifier is useful in particular when multiclass classification. The previous subsection, we present the probability estimation system is implemented as shown in Fig. 5, and it is widely tested under various physical environments. We present some classification results in Table 1 when 9 feature vectors are combined.

The proposed probabilistic method is compared with other methods such as the standard SVM with voting (the largest number of recommendation) and previous probabilistic method [5]. As listed in Table 1, the proposed method reduces the misclassified cases for almost all age classes except for class 3. Remember that class 3 includes

$$ c^* = \arg \max_c \sum_{k=1}^{M} P_{k|c}, \quad k = 1, \ldots, M $$

where $M$ is the number of feature vector, $P_{k|c}$ is the probability of class $c$ in $k^{th}$ feature vector.

3. Experimental Results

The proposed probabilistic classifier is evaluated for age group classification and its performance is verified using our own Korean Face DB that has 473 facial images of different persons with diverse ages. The DB contains natural facial images taken in daily life situations such as ‘face expression of laugh’, ‘wear hat’ and ‘wear glasses’. The face images are divided into 4 age groups that are 0~19, 20~39, 40~59 and over 60. We conduct 4 fold cross validation test with sorted ages. The Bio-Inspired Features (BIF) [8] and Histogram of Gradient (HOG) [9] features are extracted from 9 local facial regions as shown in Fig. 4. The age group estimation system is implemented as shown in Fig. 5, and it is widely tested under various physical environments. We present some classification results in Table 1 when 9 feature vectors are combined.

The proposed probabilistic method is compared with other methods such as the standard SVM with voting (the largest number of recommendation) and previous probabilistic method [5]. As listed in Table 1, the proposed method reduces the misclassified cases for almost all age classes except for class 3. Remember that class 3 includes

2.3 Application to Age Group Classification

In previous subsection, we present the probability estimation method using SVM for multiclass classification. The probabilistic classifier is useful in particular when multiclass features are classified individually and combined for an overall decision.

In this subsection, we present how to apply the probabilistic classifier to the actual classification problem, and its practical usefulness is demonstrated in the experimental results. As shown in Fig. 4, 9 different local features are extracted for age group estimation. Each feature is first probabilistically classified individually. In other words, we obtain 9 different weak classification results expressed by probability for each class. Then, they are merged for final classification. Generally expressing, the optimal classification is given by,

$$ P_i(C_i|X) = \frac{\sum_{j=1, j \neq i}^{N} f_{ij}(d_{ij})}{\sum_{k=1}^{N} \sum_{j=1, j \neq k}^{N} f_{kj}(d_{kj})}, \quad i = 1, 2, \ldots, N $$

where $f_{ij}(d_{ij})$ is the classification function.

By summing the probabilities of all pairwise classifications, we estimate the probability for multiclass classification. Since the sum of all class probabilities should be unity, the class posterior probability is normalized as shown in (3). The key feature of the proposed method is to use its own distribution for each class (Fig. 3 (a)), unlike the conventional method as shown in Fig. 3 (b). Its benefits can be described as follows. First, we can remove the effect of the opposite class distribution and find more accurate best-fit function by considering the sign of distances. As shown in Fig. 3, finding best-fit function for each class respectively brings more accurately estimated function. Second, we reduce misclassification caused by score-sensitive SVM outputs. The previous method estimates probability of class $C_i$ and $C_j$ ($P_{ij}$, $P_{ji}$) using sigmoid function from training scores of both classes (Fig. 3 (b)). In the score-sensitive case, that is the unknown observation is just located near hyperplane, the classification probabilities for both class $C_i$ and $C_j$ are equal to each other. This works well for binary classification. However, its extension to multiclass classification using “pairwise coupling” strategy results in less differentiation of probability among classes. On the other hand, the proposed method estimates probability with degree of confidence. The proposed method estimate higher probability only for the case with high confidence. In score-sensitive case, the method estimates little probability. This reflects confidence of every pairwise class for an overall decision. In practical applications, there are many misclassified unknown observation located near hyperplane in feature space, and it is important to decrease these misclassification. Therefore, we use different probabilistic functions for pairwise class while the conventional method uses a single function.

![Facial regions used for feature extraction.](image)

![Actual implementation of age group classification with the proposed probabilistic SVM method.](image)
Table 1: Classification accuracy.

| Ground Truth Class | C1 | C2 | C3 | C4 |
|--------------------|----|----|----|----|
| Estimated Class    |    |    |    |    |
| C1                 | 88.89 | 8.57 | 4.17 | 0.00 |
| C2                 | 11.11 | 82.86 | 29.17 | 0.00 |
| C3                 | 0.00  | 2.86  | 45.83 | 12.50 |
| C4                 | 0.00  | 5.71  | 20.83 | 87.50 |

Total : 78.15

Fig. 6: Estimated mean probability of each class for given ground truth classes; (a) previous method [5] (b) proposed method.

ages of 40–59, and it actually shows less common noticeable characteristics relatively, compared to other age groups. It contains much more diverse sample images, and that is a primary reason for low accuracy. Next, individual local features are combined into a single multidimensional set of features, which is applied to the conventional and proposed methods. As listed in Table 1 (d), the classification results are identical between the standard SVM and the probabilistic SVMs because the probabilistic SVMs output a class with the largest probability, which actually corresponds to a class with the largest score. However, the classification result is less accurate than the combination of weak classifiers.

Figure 6 shows the comparisons of the estimated mean probability of each class for a given ground truth. The proposed method shows more outstanding probability difference between age classes.

That means that we can estimate the probability with high confidence. These practical experimental results show the contribution to the improvement of classification accuracy.

4. Conclusion

In this letter, we proposed a multiclass probabilistic classification method based on SVM. We use the distances of the unknown observation from each pairwise hyperplane to estimate the probability of the classes. Most of previous age group classification research uses the deterministic SVM. This motivates us to introduce the probability in age group classification. We apply the proposed probabilistic method to 4 class age group classification and confirm more accurate estimation results, compared to the deterministic SVM. The proposed method reduces misclassification caused by the effect of score-sensitive classifications and shows superior classification capability in terms of both accuracy and robustness.

Acknowledgments

This research was funded and supported by Samsung Electronics Co., Ltd.

This research was supported by the MSIP (Ministry of Science, ICT and Future Planning), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2015-H8501-15-1017) supervised by the IITP (Institute for Information & communications Technology Promotion).

References

[1] C. Burges, “A tutorial on support vector machines for pattern recognition,” Data Mining and Knowledge Discovery, vol.2, no.2, pp.121–167, June 1998
[2] J.D. Txia and C.L. Huang, “Age estimation using AAM and local facial features,” Proc. 5th International Conference on Intelligent Information Hiding and Multimedia Signal Processing (IIH-MSP), pp.885–888, Sept. 2009.

[3] L. Li, J. Liu, and J. Cheng, “Age-group classification of facial images,” 11th International Conference on Machine Learning and Applications (ICMLA), vol.1, pp.693–696, Dec. 2012.

[4] J.-S. Bae and J.-O. Kim, “Age group classification using geometry features,” Proc. International Technical Conference on Circuits/Systems Computers and Communications (ITC-CSCC), pp.554–555, July 2013.

[5] J. Platt, “Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods,” Advances in Large Margin Classifiers, vol.10, no.3, pp.61–74, Oct. 1999.

[6] M. Kugler, S. Kuroyagi, A.S. Nugroho, and A. Iwata, “CombNet-III: A support vector machine based large scale classifier with probabilistic framework,” IEICE Trans. Inf. & Syst., vol.E89-D, no.9, pp.2533–2541, Sept. 2006.

[7] M. Parandehgheibi, “Probabilistic classification using fuzzy support vector machines,” Proc. 6th INFORMS Workshop on Data Mining and Health Informatics (DM-HI), Nov. 2011.

[8] G. Guo, G. Mu, Y. Fu, and T.S. Huang, “Human age estimation using bio-inspired features,” IEEE Computer Vision and Pattern Recognition (CVPR), pp.112–119, Florida, USA, June 2009.

[9] F. Samadi, “Human age-group estimation based on ANFIS using the HOG and LBP features,” Electrical and Electronics Engineering: An International Journal (ELEIJ), vol.2, no.1, pp.21–29, Feb. 2013.