Image Classification Using the Max-Min Posterior Pseudo-Probabilities Method

Xiabi Liu*, Yunde Jia, Xuefeng Chen, Yuan Deng, and Hui Fu

Beijing Laboratory of Intelligent Information Technology, School of Computer Science and Technology, Beijing Institute of Technology, Beijing 100081, China

***Abstract*—**This paper proposes a novel approach to Bayesian pattern classification and explores it for classifying images. A notion of posterior pseudo-probability is introduced to imitate posterior probability. Classification decisions are made upon the values of posterior pseudo-probabilities which are computed from class-conditional densities by an advised family of functions. We further present a discriminative learning algorithm called Max-Min posterior Pseudo-probabilities (MMP) to learn unknown parameters in the mapping function between class-conditional densities and posterior pseudo-probabilities. The main idea behind the MMP learning is to optimize the classifier performance through maximizing posterior pseudo-probabilities for each class and its positive samples, while minimizing those for each class and its negative samples. The proposed MMP approach to Bayesian pattern classification was tested in two tasks of image classification, including text extraction and content-based image retrieval. In the experiments, the MMP method was compared with the maximum likelihood based method, the minimum classification error method, and support vector machines. The experimental results show the effectiveness of our approach.

***Index Terms*—** Discriminative learning; Bayesian pattern classification; Bayesian classifiers; Maximum Likelihood (ML); Expectation-Maximization (EM) algorithm, Minimum Classification Error (MCE), Support Vector Machines (SVMs)

* Corresponding Author. Tel: +86 10 68913447, Fax: +86 10 86343158, E-mail: liuxiabi@bit.edu.cn

Ⅰ. INTRODUCTION

It is important for a statistical pattern classifier to learn representative class information from the samples of classes. Learning approaches to statistical pattern classification can be divided into two categories: generative learning and discriminative learning. They are differentiated by their criterions to evaluate learning results. In generative learning algorithms, such as in classical Maximum Likelihood (ML) based algorithms, the first concern is the fit of the class model to observed data. The discrimination between classes is realized indirectly but guaranteed by Bayesian decision theory. Because of the insufficiency of the training data or noises in the training data, the class models estimated by generative learning algorithms often deviate from satisfactory ones, which lead to unsatisfactory classifiers. In order to solve this problem, discriminative learning algorithms are introduced to directly consider the discrimination between classes in the training phase. They focus on the difference between classes, instead of only the distribution of a single class. In recent applications to a wide range of classification tasks, discriminative learning algorithms demonstrate significant better performance over generative counterparts [1-8], or are used to enhance generative learning based classifiers [9-11].

Commonly used discriminative learning approaches include Support Vector Machines (SVMs) [12], Minimum Classification Error (MCE) methods [13], Maximum Mutual Information (MMI) methods [14], and Neural Networks. In SVMs, the upper bound of the generalization error is minimized through maximizing the margin between the separating hyper-plane and the training data. MMI methods are intended to minimize amount of uncertainty about classification through maximizing the mutual information between class models and the training data, while MCE methods aim to minimize the error rate on the training data. Recently, people developed some other discriminative criterions, such as

Minimum Word or Phone Error (MWE/MPE) [15], Figure Of Merit (FOM) [16], and Margin based ones [6, 17].

In this paper, we propose a novel approach to Bayesian pattern classification and its related discriminative learning algorithm. Through investigating Bayes formula from a new point of view, we introduce a notion of posterior pseudo-probability as the imitation of a posterior probability. A family of functions is then advised to compute the values of posterior pseudo-probabilities from class-conditional densities. We further present a discriminative learning algorithm called Max-Min posterior Pseudo-probabilities or MMP for short to learn parameters in the mapping function between class-conditional densities and posterior pseudo-probabilities. In the MMP learning, the optimal classifier is considered to be achieved if the posterior pseudo-probabilities for each class and its positive samples are measured as 1, while those for each class and its negative samples are measured as 0. In light of this idea, the MMP learning objective is defined and optimized using the gradient descent algorithm.

To perform classification tasks, a mapping function between class-conditional densities and posterior pseudo-probabilities is assigned to each class. The function parameters are estimated from positive and negative samples of the class using the MMP learning algorithm. Given an input pattern, the corresponding posterior pseudo-probability for each class is computed. The input pattern is then classified into the class with maximum posterior pseudo-probability or rejected as being unrecognized if the maximum posterior pseudo-probability is below a threshold.

We apply the proposed MMP approach to Bayesian pattern classification to content-based image retrieval. An early version of the MMP, called Maximum-Minimum Similarity or MMS for short [18], has also been applied to text extraction In the experiments, the performance of the MMP approach is

compared with that of the baseline Expectation-Maximization (EM) algorithm in a ML setting, the MCE method, and the SVM. The experimental results show that our approach is promising and effective.

The rest of this paper is organized as follows. Section 2 introduces the notion of posterior pseudo-probability and its measure functions. Section 3 presents the MMP learning algorithm. Two applications of the MMP method in image classification and corresponding experimental results are reported in section 4. We discuss conclusions and future works in Section 5

Ⅱ. BAYESIAN PATTERN CLASSIFICATION USING POSTERIOR PSEUDO-PROBABILITIES

Given a feature vector $\mathbf{x}$, a finite set of classes $\{\omega_1,\cdots,\omega_n\}$. Let $P(\omega_i)$, $p(\mathbf{x}|\omega_i)$, $P(\omega_i|\mathbf{x})$ be the prior probability, the class-conditional probability density function, and the posterior probability, respectively. Bayes classification rule for minimizing the probability of error is to classify $\mathbf{x}$ into the class ω^* with maximum posterior probability, i.e.

$$\omega^* = \arg\max_{\omega_i} P(\omega_i|\mathbf{x}). \tag{1}$$

Based on Bayes formula

$$P(\omega_i|\mathbf{x}) = \frac{p(\mathbf{x}|\omega_i)P(\omega_i)}{p(\mathbf{x})}, \tag{2}$$

We have

$$\frac{P(\omega_i|\mathbf{x})}{P(\omega_j|\mathbf{x})} = \frac{p(\mathbf{x}|\omega_i)P(\omega_i)}{p(\mathbf{x}|\omega_j)P(\omega_j)}. \tag{3}$$

Thus, the completely equivalent and actually used decision rule is given by

$$\omega^* = \arg\max_{\omega_i} p(\mathbf{x}|\omega_i)P(\omega_i). \tag{4}$$

Now we investigate Bayes formula in Eq. 2 from another point of view. Given two values $\mathbf{x}$ and $\mathbf{x}'$ of feature vector, we can obtain

$$\frac{P(\omega_i|\mathbf{x})}{P(\omega_i|\mathbf{x}')} = \frac{p(\mathbf{x}|\omega_i)p(\mathbf{x}')}{p(\mathbf{x}'|\omega_i)p(\mathbf{x})} \tag{5}$$

according to Eq. 2. It should be noted that Eq. 5 is totally different with Eq. 3. In Eq. 5, two observations and a single class are involved. Oppositely, a single observation and two classes are involved in Eq. 3.

If we consider that $\mathbf{x}$ is evenly measured in the whole feature space, instead of only in the sub-space constrained by the set of classes, then it is reasonable to assume that $\mathbf{x}$ is distributed uniformly. Accordingly, $p(\mathbf{x}) = p(\mathbf{x}')$ and Eq. 5 can be simplified as

$$\frac{P(\omega_i|\mathbf{x})}{P(\omega_i|\mathbf{x}')} = \frac{p(\mathbf{x}|\omega_i)}{p(\mathbf{x}'|\omega_i)} \tag{6}$$

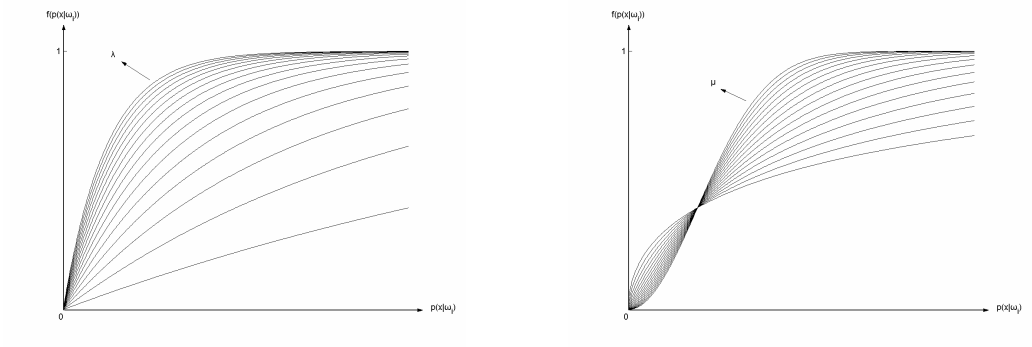
which means

$$P(\omega_i|\mathbf{x}) \propto p(\mathbf{x}|\omega_i). \tag{7}$$

According to Eq. 7, we can imitate $P(\omega_i|\mathbf{x})$ through embedding $p(\mathbf{x}|\omega_i)$ in a smooth, monotonically increasing function which takes value in $[0,1]$. We call the values of this kind of functions as posterior pseudo-probabilities. Let λ and μ are positive numbers, then the function

$$f\left(p(\mathbf{x}|\omega_i)\right) = 1 - \exp\left(-\lambda p^{\mu}(\mathbf{x}|\omega_i)\right) \tag{8}$$

is a smooth, monotonically increasing function of $p(\mathbf{x}|\omega_i)$, and $f(0) = 0$ and $f(+\infty) = 1$. So it is chosen to compute posterior pseudo-probabilities in this paper. Fig. 1 shows the family of $f\left(p(\mathbf{x}|\omega_i)\right)$ generated by varying λ or μ.



(a) (b)

Fig. 1. The family of posterior pseudo-probability measure function (Eq. 8): (a) generated by varying λ; (b) generated by varying μ

The pattern classification method based on posterior pseudo-probabilities includes two stages. In the learning stage, the posterior pseudo-probability measure function (Eq. 8) of each class is learned from observed data using the MMP learning algorithm which is described in the next section. In the classification stage, the posterior pseudo-probability for each class and the input pattern is measured. The input pattern is then classified into the class with maximum posterior pseudo-probability or rejected as being unrecognized if the maximum posterior pseudo-probability is below a threshold. An intuitive and reasonable threshold for making rejection decision is 0.5. This classification rule can be represented as

$$\omega^* = \begin{cases} \arg\max\limits_{\omega_i} f\left(p\left(\mathbf{x}|\omega_i\right)\right), \max f\left(p\left(\mathbf{x}|\omega_i\right)\right) > 0.5 \\ \text{unrecognized}, \max f\left(p\left(\mathbf{x}|\omega_i\right)\right) \le 0.5 \end{cases} \tag{9}$$

Compared with traditional Bayesian classification rule (Eq. 4), the advantage of posterior pseudo-probabilities based classification rule (Eq. 9) is that the value for making decision is in $[0,1]$, so it is a natural similarity measure and is useful for (1) making rejection decision, (2) combining classifiers, (3) assessing the performance of a classifier in a much more accurate way than that of counting the number of patterns classified correctly [19].

III. MAX-MIN POSTERIOR PSEUDO-PROBABILITIES LEARNING

In this section, we present an algorithm for learning posterior pseudo-probability measure function of each class from observed data. For simplifying the problem, we assume $p(\mathbf{x}|\omega_i)$ in Eq. 8 is of some known form, and only a few parameters are unknown. Consequently, the task is to estimate parameters in Eq. 8, including λ, μ and those in $p(\mathbf{x}|\omega_i)$.

A. Learning Criterion

We can imagine a perfect Bayes classifier in which given any pattern, the posterior probability for its true class is measured as 1, and those for its false classes as 0. This implies that the classification performance of a posterior pseudo-probabilities based classifier can be optimized by producing the posterior pseudo-probability measure function of each class in order that the posterior pseudo-probabilities measured for positive samples of this class are maximized towards 1, while those for its negative samples are minimized towards 0. We call this learning idea as Max-Min posterior Pseudo-probabilities or MMP for short.

More formally, let $f(\mathbf{x};\boldsymbol{\Lambda})$ be the posterior pseudo-probability measure function of a class, where $\boldsymbol{\Lambda}$ denote the set of unknown parameters in it. Let $\hat{\mathbf{x}}_i$ be the feature vector of arbitrary positive sample of the class, $\bar{\mathbf{x}}_i$ be the feature vector of arbitrary negative sample of the class, m and n be the number of positive and negative samples of the class, respectively. According to the idea above of the MMP learning, the objective function for estimating parameters is designed as

$$F(\boldsymbol{\Lambda}) = \frac{1}{m}\sum_{i=1}^{m}\left[f(\hat{\mathbf{x}}_i;\boldsymbol{\Lambda})-1\right]^2 + \frac{1}{n}\sum_{i=1}^{n}\left[f(\bar{\mathbf{x}}_i;\boldsymbol{\Lambda})\right]^2 . \tag{10}$$

$F(\boldsymbol{\Lambda})=0$ means the perfect classification performance on the training data. Consequently, we can obtain the optimum parameter set $\boldsymbol{\Lambda}^*$ of the posterior pseudo-probability measure function by minimizing $F(\boldsymbol{\Lambda})$:

$$\mathbf{\Lambda}^* = \arg\min_{\mathbf{\Lambda}} F(\mathbf{\Lambda}). \tag{11}$$

B. Optimization Method

In this paper, the gradient descent algorithm is applied to optimize the parameter set of each posterior pseudo-probability measure function according to Eq. 11. In fact, the following iterative equation is used to update the parameters:

$$\mathbf{\Lambda}_{t+1} = \mathbf{\Lambda}_t - \alpha_t \nabla F(\mathbf{\Lambda}_t), \tag{12}$$

where $\mathbf{\Lambda}_t$ and α_t is the parameter set and the step size in the t-th iteration, respectively, $\nabla F(\mathbf{\Lambda}_t)$ is the partial derivatives of $F(\mathbf{\Lambda})$ with respect to all the parameters in $\mathbf{\Lambda}_t$. Let ψ denote arbitrary parameter in $\mathbf{\Lambda}$, then we have

$$\frac{\partial F}{\partial \psi} = \frac{2}{m}\sum_{i=1}^{m}\left(f(\hat{\mathbf{x}}_i;\mathbf{\Lambda})-1\right)\frac{\partial f(\hat{\mathbf{x}}_i;\mathbf{\Lambda})}{\partial \psi} + \frac{2}{n}\sum_{i=1}^{n} f(\overline{\mathbf{x}}_i;\mathbf{\Lambda})\frac{\partial f(\overline{\mathbf{x}}_i;\mathbf{\Lambda})}{\partial \psi}. \tag{13}$$

In Eq. 13, $\frac{\partial f(\hat{\mathbf{x}}_i;\mathbf{\Lambda})}{\partial \psi}$ and $\frac{\partial f(\overline{\mathbf{x}}_i;\mathbf{\Lambda})}{\partial \psi}$ depend on $f(\mathbf{x};\mathbf{\Lambda})$ and ψ, and have to be decided in the applications.

According to Eqs. 12-13, the MMP algorithm for learning the posterior pseudo-probability measure function of each class is described as follows. The whole procedure of the MMP learning is to perform this algorithm one by one for all classes.

Step1. Compute the partial derivative of $F(\mathbf{\Lambda})$ with respect to each parameter using Eqs. 13.

Step2. Compute the step size α_t using the improved 0.618 method [20].

Step3. Update the parameters using Eq. 12.

Step4. Repeat Step 1 to Step 3 until convergence or the preset maximum number of iterations is reached.

Let ε be an infinitesimal, the convergence condition is

$$\|g_t\| = \left(\sum\left(\frac{\partial f(\mathbf{x};\mathbf{\Lambda})}{\partial \psi}\right)^2\right)^{1/2} \le \varepsilon \quad . \tag{14}$$

Ⅳ. CASE STUDIES

The proposed MMP approach to Bayesian pattern classification is applied to text extraction and content-based image retrieval (CBIR). In both applications, the form of class-conditional probability density function $p(\mathbf{x}|\omega_i)$ in Eq. 8 is assumed to be the Gaussian Mixture Model (GMM). Let K be the number of Gaussian components in GMM, w_k, $\boldsymbol{\mu}_\mathbf{k}$ and $\boldsymbol{\Sigma}_\mathbf{k}$ respectively be the weight, the mean, and the covariance matrix of the k-th Gaussian component, $\sum_{k=1}^{K} w_k = 1$, then we have

$$p(\mathbf{x}|\omega_i) = \sum_{k=1}^{K} w_k N(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), \tag{15}$$

where

$$\begin{aligned} &N(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \\ &= (2\pi)^{-\frac{d}{2}} |\boldsymbol{\Sigma}_k|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1}(\mathbf{x}-\boldsymbol{\mu}_k)\right). \end{aligned} \tag{16}$$

So the set of unknown parameters in the posterior pseudo-probability measure function of each class is

$$\boldsymbol{\Lambda} = \{\lambda, \mu, w_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}, k = 1, \cdots, K. \tag{17}$$

In all the following experiments, we firstly used the EM algorithm on positive samples of the class to get the Maximum Likelihood Estimation (MLE) of parameters in GMM, and set λ and μ through careful experiments. Then the MMP learning algorithm was used on all the samples including positive samples and negative samples to revise the initial parameters obtained by the EM algorithm.

A. Text Extraction

In the application to multilingual text extraction, an early version of MMP, called Maximum-Minimum Similarity or MMS for short, is used to discriminate character regions from non-character regions in images. In the experiments, our text extraction approach with the MMS learning achieved the recognition rate of 93.6% for the test data set, which is better than not only 81.1% coming from the baseline EM algorithm but also 82.2% coming from the MCE learning.

The details on the MMS based text extraction approach and corresponding experimental results can be found in our previous paper [18]

B. Content-Based Image Retrieval

In the application to CBIR, we consider the problem of retrieving images by their semantic concepts. Each image is represented as an 80-D low-level feature vector which consists of 9-D color moments and 71-D Gabor texture features. The low-level feature vectors of images are linked with their high-level semantics concepts using the MMP classification method. In the learning stage, a posterior pseudo-probability function is learned for each semantic concept using the MMP learning algrothim. In the stage of image retrieval, the images in the database are classified into relevant or irrelevant to the query concept according to corresponding posterior pseudo-probabilities.

The experiments involve 5000 images from Corel database [21]. These images are divided into 50 categories, each of which includes 100 images. The category names of images are thought to be their semantic concepts. We used 50% of 5000 images to learn posterior pseudo-probability measure functions of all categories and conducted two kinds of experiments of querying by concepts on the rest images. In both kinds of experiments, all test images were sorted in descending order of posterior pseudo-probabilities measured for the query concept and them. Then 50 top rank images were retrieved as results in the first kind of experiment. So the precision rate is equal with the recall rate in this kind of experiments. But in the second kind of experiments, we retrieved all the images for which the posterior pseudo-probabilities are larger than 0.5. The experimental results were obtained using the estimated parameters by the baseline EM algorithm and the MMP learning algorithm, respectively. The maximum iteration number in the MMP learning is set to 200.

Furthermore, the SVM was also tested in the same experiments, where the distance between the sample and the decision boundary is used as the prediction confidence as in other SVM based image retrieval methods [22-23]. Therefore, all test images were sorted in descending order of the distances between them and the decision boundary corresponding with the query concept. Then 50 top rank images were

retrieved as results in the first kind of experiments, while all the images classified as the query concept by the SVM were returned in the second kind of experiments.

The experimental performance comparison of the EM, the MMP, and the SVM are listed in Table 1, where 'PR' means the Precision Rate, 'RR' means the Recall Rate, 'NRI' means the average Number of Retrieved Images, the superscript '1' and '2' respectively mean the first and second kind of experiments.

It should be noted that the EM algorithm and the SVM were implemented using Torch machine learning library [24].

TABLE Ⅰ

PERFORMANCE COMPARISON OF EM, MMP, AND SVM IN CBIR EXPERIMENTS

Learning algorithms	PR^1	PR^2	RR^2	NRI^2
EM	19.12%	6.95%	79.2%	1064.8
MMP	27%	19.90%	44.80%	174.9
SVM	8.84%	33.06%	8.12%	10.1

In the second kind of experiments with the MMP learning, the best result of closed test is obtained from the concept 'Easter Egg'. The corresponding PR, RR, and NRI are 100%, 86%, and 43, respectively. And the worst result came from the concept 'New York City', where PR, RR, and NRI are 4.89%, 22%, and 225, respectively. The corresponding results obtained by the SVM are: 100% (PR), 22% (RR), 11 (NRI) for the concept 'Easter Egg', and 0 (PR), 0 (PR), 2 (NRI) for the concept 'New York City'.

Fig. 2-3 show 20 top rank images with corresponding posterior pseudo-probabilities for the concept 'Easter Egg' and 'New York City', respectively, where the symbol '√' indicates relevant images and '×' indicates irrelevant image. All the images shown in Fig. 2 are expected results by the user. In fact, all 43 images returned are relevant images in this case. However, most of images shown in Fig. 3 are irrelevant to the query concept 'New York City'. The reason behind the huge difference of performance

on the two concepts exists in the feature stability of corresponding images. As shown in Fig. 2, there is a distinctly common feature among the images with the concept ‘Easter Egg’ in the Corel database, which is captured by the feature vector used in this paper. Oppositely, the images with the concept ‘New York City’ in the Corel database are so diverse that it is hard to extract the common feature from them. This difficulty is shown in the example images with the concept ‘New York City’ in Fig. 4.

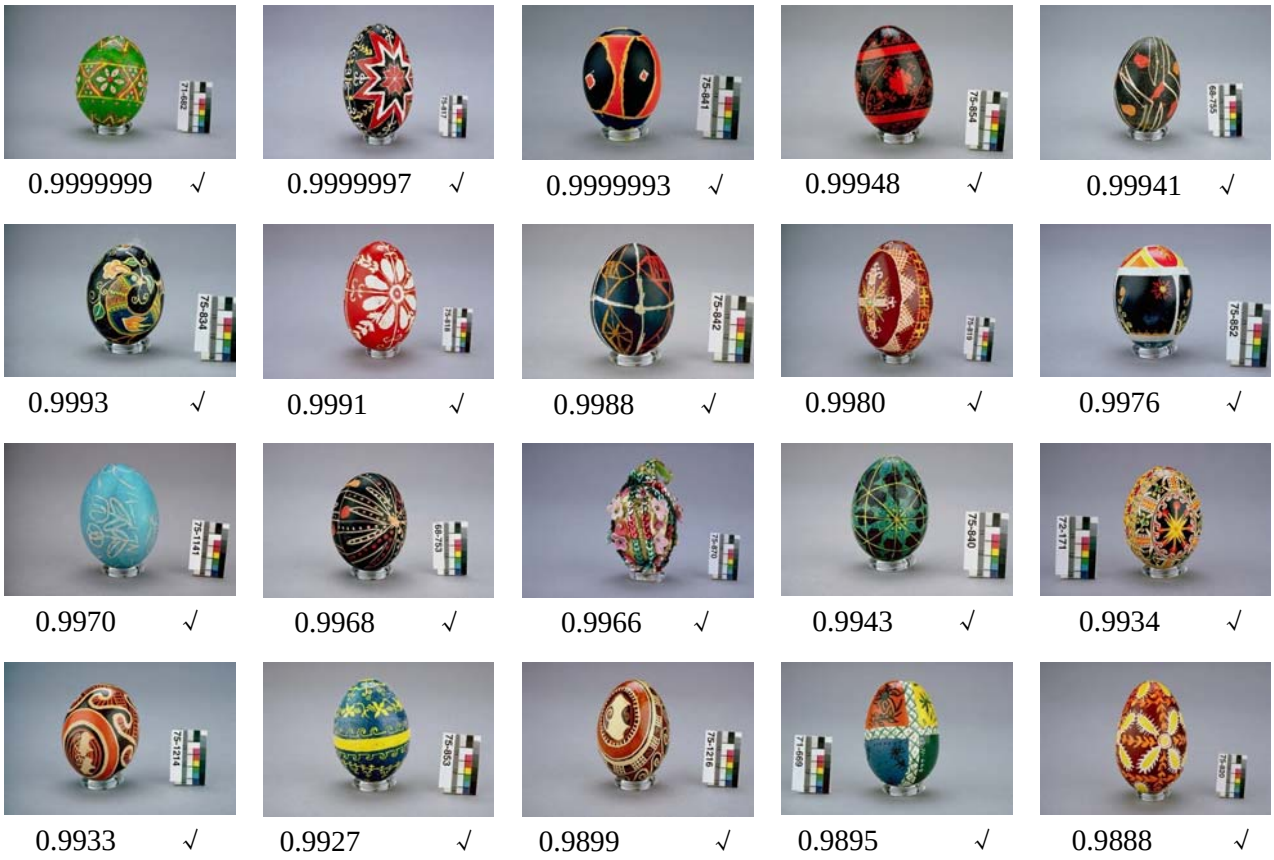


Fig. 2. 20 top rank images retrieved and corresponding posterior pseudo-probabilities for the concept ‘Easter Egg’ in the closed test with the MMP learning.

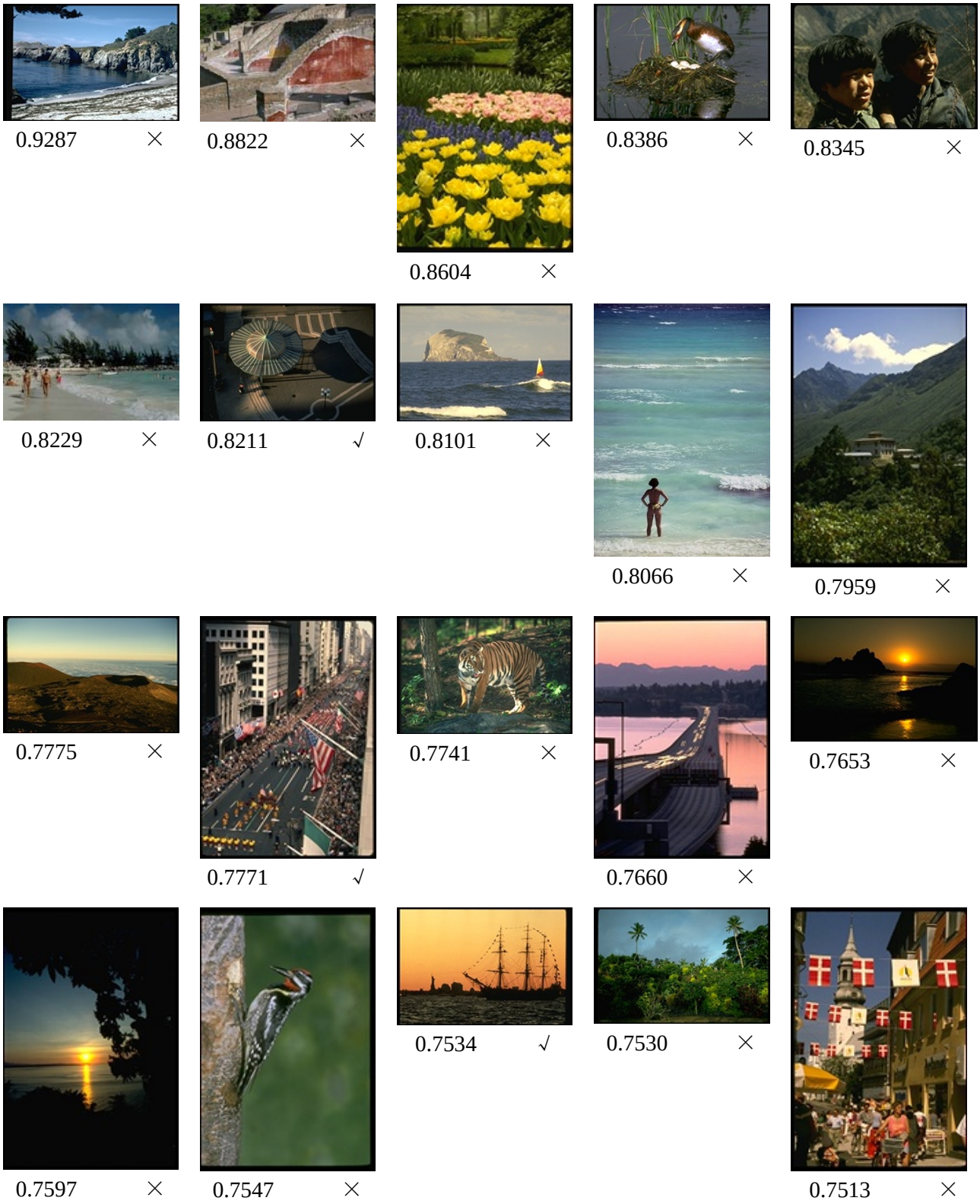


Fig. 3. 20 top rank images retrieved and corresponding posterior pseudo-probabilities for the concept 'New York City' in the closed test with the MMP learning.

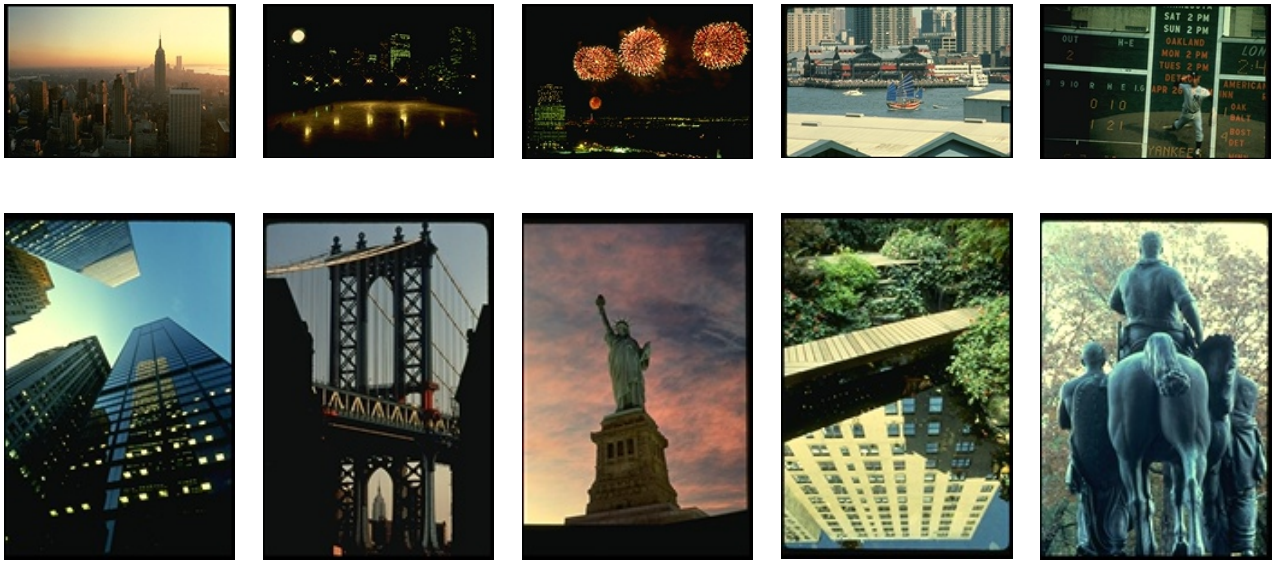


Fig. 4. The example images with the concept ‘New York City’ in the Corel database

Ⅴ. CONCLUSIONS

In this paper, we have proposed a novel approach to Bayesian pattern classification, which is called Max-Min posterior Pseudo-probability or MMP for short. Two main contributions of this paper are summarized as follows.

(1) A notion of posterior pseudo-probability is introduced as the imitation of a posterior probability. The values of posterior pseudo-probabilities are measured from class-conditional densities by an advised family of functions and used to make classification decisions.

(2) A MMP learning algorithm is presented to learn unknown parameters in the posterior pseudo-probability measure function of each class from observed data.

Compared with traditional Bayesian classification rule, the advantage of MMP method is that posterior psudo-probabilities take values in $[0,1]$, so it is a natural similarity measure and is useful for (1) making rejection decision, (2) combining classifiers, (3) assessing the performance of a classifier in a much more accurate way than that of counting the number of patterns classified correctly.

We have tested the proposed MMP classification method in two tasks of image classification, including text extraction and content-based image retrieval. In the experiments, the MMP method behaved better than not only the classical generative learning method of EM algorithm in a ML setting, but also the commonly used discriminative learning methods of the MCE and the SVM.

There are several open problems in current MMP method, including (1) the possibility and the effectiveness of using other forms of posterior pseudo-probability measure functions; (2) the use of more effective and more efficient optimization method; (3) the analysis of the convergence and the speed of the MMP learning algrothm.

REFERENCES

[1] Rong Yan, Jian Zhang, Jie Yang, and Alexander G. Hauptmann, "A discriminative learning framework with pairwise constraints for video object classification," *IEEE Trans. Pattern Analysis and Machine Learning*, vol. 28, no. 4, pp. 578-593, April. 2006.

[2] Cuzmán Santafé, Jose A. Lozno, and Pedro Larrañaga, "Discriminative learning of Bayesian network classifiers via the TM algorithm," in *Proc. 8th Euro. Conf. Symbolic and Quantitative Approaches to Reasoning with Uncertainty* (*ECSQARU 2005)*, Catalonia, Spain, 2005, pp. 148-160.

[3] Minyoung Kim and Vladimir Pavlovic, "Discriminative learning of mixture of Bayesian network classifiers for sequence classification," in *Proc. of the 2006 IEEE Computer Society Conf. on Computer Vision and Pattern Recognition (CVPR'06)*, vol. 1, New York, 2006, pp. 268-275.

[4] Minyoung Kim and Vladimir Pavlovic, "A recursive method for discriminative mixture learning," in *Porc. of the 24th International Conference on Machine Learning*, Corvallis, OR, 2007, pp. 409-416.

[5] Erik McDermott, Timothy J. Hazen, Jonathan Le Roux, Atsushi Nakamura, and Shigeru Katagiri, "Discriminative training for large-vocabulary speech recognition using minimum classification error," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 15, no. 1, pp. 203-223, Jan. 2007.

[6] Jinyu Li, Ming Yuan, and Chin-Hui Lee, "Approximate test risk bound minimization through soft margin estimation," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 15, no. 8, pp. 2393-2404, Nov. 2007.

[7] Alain Biem, "Minimum classification error training for online handwriting recognition," *IEEE Trans. on Pattern Recognition and Machine Intelligence*, vol. 28, no. 7, pp. 1041-1051, Jul. 2006.

[8] Oksana Yakhnenko, Adrian Silvescu, and Vasant Honavar, "Discriminatively trained Markov model for sequence classification," in *Proc. of the Fifth International Conf. on Data Mining (ICDM'05)*, Houston, Texas, 2005, pp. 498-505.

[9] Alex Holub and Pietro Perona, "A discriminative framework for modeling object classes," in *Proc. of the 2005 IEEE Computer Society Conf. on Computer Vision and Pattern Recognition (CVPR'05)*, San Diego, CA, 2005, pp.664-671.

[10] Yi Li, Linda G. Shapiro, and Jeff A. Bilmes, "A generative/discriminative learning algorithm for image classification," in *Proc. of the Tenth IEEE International Conference on Computer Vision (ICCV'05)*, vol. 2, Beijing, 2005, pp. 1605-1612.

[11] Yanmin Sun, Andrew K. C. Wong, and Yang Wang, "Generative and discriminative learning by CL-Net," *IEEE Trans. Systems, Man, and Cybernetics-Part B: Cybernetic*, vol. 37, no. 4, pp. 1-8, Aug. 2007.

[12] V. N. Vapnik. *The nature of statistical learning theory*. New York: Springer, 1995.

[13] Biing-Hwang Juang and Shigeru Katagiri, "Discriminative learning for minimum error classification," *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. 40, no. 12, pp. 3043-3054, Dec. 1992.

[14] Lalit R. Bahl, Peter F. Brown, Peter V. de Souza and Robert L. Mercer, "Maximum mutual information estimation of hidden Markov model parameters for speech recognition," In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, New York, NY, USA, 1986, pp. 49-52.

[15] D. Povey and P. C. Woodland, “Minimum phone error and i-smoothing for improved discriminative training,” In *Proceedings of 2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. I , Orlando, FL, USA, May 2002, pp. 105-108.

[16] Xiaohan Li, Eric Chang and Bei-qian Dai, “Improving speaker verification with figure of merit training,” In *Proceedings of 2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. I , Orlando, FL, USA, May 2002, pp. 693-696.

[17] H. Jiang, X. Li, and C. Liu, “Large margin hidden Markov models for speech recognition,” *IEEE Trans. Audio, Speech, and Language Processing*, vol. 14, no. 5, pp. 1584-1595, Sept. 2006.

[18] Xiabi Liu, Hui Fu, and Yunde Jia, “Gaussian mixture modeling and learning of neighboring characters for multilingual text extraction in images,” *Pattern Recognition*, vol. 41, no. 2, pp. 484-493, Feb. 2008.

[19] G. Lugosi, and M. Pawlak, “On the posterior-probability estimate of the error rate of nonparametric classification rules,” *IEEE Trans. Information Theory*, vol. 40, no. 2, pp. 475-481, Mar. 1994.

[20] Yaxiang Yuan and Wenyu Sun, *Optimization Theory and Methods (in Chinese)*. Beijing: Science Press, 2003

[21] Corel Image Database. Available: http://www.corel.com

[22] Jian Cheng and Kongqiao Wang, “Active learning for image retrieval with Co-SVM,” *Pattern Recognition*, vol. 40, no. 1, pp. 330-334, Jan. 2006.

[23] Jing Li, Nigel Allinson, Dacheng Tao, and Xuelong Li. “Multitraining support vector machine for image retrieval,” *IEEE Trans. Image Processing*, vol. 15, no. 11, pp. 3597-3601, Nov. 2006.

[24] Torch Machine Learning Library. Available: http://www.torch.ch