FACE RECOGNITION USING IMPROVED PAIRWISE COUPLING SUPPORT VECTOR MACHINES

Zeyu LI [1,*] , *Shiwei TANG* [1,2]

[1] National Laboratory on Machine Perception and Center for Information Science,
Peking University, Beijing, 100871
[2] Department of Computer Science and Technology, Peking University, Beijing, 100871

ABSTRACT

In this paper, a novel structure is proposed to tackle multi-class classification problem. For a K-class classification task, an array of K optimal pairwise coupling classifier (*O-PWC*) is constructed, each of which is the most reliable and optimal for the corresponding class in the sense of cross entropy or square error. The final decision will be got through combining the results of these K *O-PWC*s. The accuracy rate is improved while the computational cost will not increase too much. This algorithm is applied to face recognition on Cambridge ORL face database, the experimental results reveal our method is effective and efficient.

1. INTRODUCTION

In many real world applications, such as face recognition, text categorization, handwritten digital recognition, and so on, a multi-class classification problem has to be solved. One method is to establish a unified hyperplane to discriminate all classes at once directly [8]. More popular and applicable method is to reduce a multi-class problem to a set of binary classification problems rather than to construct a decision function for all classes [2].

There are different strategies to decompose a multi-class problem into a number of binary classification problems. For a K-class classification problem, one method is to use *one-against-rest* [2] principle to construct K binary classifiers. Each binary classifier distinguishes one class from all the other classes. The other is so called *one-against-one* [7]. This method constructs all possible K(K-1)/2 two-class classifiers, each of which is used to discriminate two of the K classes. In this paper, we only consider the later method.

Different schemes are used to combine the results of these binary classifiers. MaxVoting strategy considers the output of each classifier as binary decision and selects the class that wins maximum votes. DAGSVM [6] constructs a Direct Acyclic Graph. Both of these methods didn’t consider the case in which the binary classifiers outputs a score whose magnitude is a measure of confidence. In Pairwise Coupling (in short, *PWC*) [10], each of the binary classifiers output a posterior probability, so called pairwise probability, for a given testing pattern. And then *PWC* couples these pairwise probabilities into a common set of posterior probabilities. This method is used widely in many fields[3][12]. Error Correct Output Codes (ECOC)[9] allows a correct classification even if a subset of binary classifiers gives wrong classification results.

However, PWC method has some drawbacks [3][4]. When a sample $\vec{x}$ is classified by one of the K(K-1)/2 classifiers, and at the same time, $\vec{x}$ doesn’t belong to both of the two involved classes of this classifier, the probabilistic measures of $\vec{x}$ to the two classes are meaningless and maybe damage the coupling output of PWC. To tackle the problem, PWC-CC method is proposed in [4]. For each pairwise classifier separating class c_i from class c_j, an additional classifier separating the two classes from the other classes will be trained. This will lead to the increment of the computational cost.

In this paper, *optimal PWC* (in short, *O-PWC*) is introduced to overcome the problem encountered by PWC. For a K-class classification problem, an array of K *O-PWC*s are constructed, each of which is optimal to the corresponding class in the sense of cross entropy or square error. Classifying a pattern equals to find the class label which corresponds to the minimal cross entropy or square error. Improved performance can be achieved while the computational cost will not increase too much.

The rest of the paper is organized as follows: In section 2, we will briefly introduce the conventional PWC method. In section 3, our algorithm is described in detail. Experimental results and conclusion will be given in section 4 and 5 respectively.

+ Supported by the Special Funds for Major State Basic Research of China, No.G1999032705.

* Corresponding author, Email: zeyul@cis.pku.edu.cn

2. PAIRWISE COUPLING METHOD

Given a set of K classes $\{c_i\}$, the probability of $\vec{x}$ belonging to class c_i, given $\vec{x}$ is in either class c_i or c_j, can be written as a pairwise probability:

$$r_{ij} = p(c_i \mid \vec{x},\ \vec{x} \in c_i \cup c_j)\ ,\ j \neq i\ .$$

Going through all K(K-1)/2 binary classifiers, a *pairwise probability matrix*, can be produced.

To couple the pairwise probability matrix into a common set of probabilities P_i, Haste and Tibshirani in [10] proposed Pairwise Coupling method. They introduced a new set of auxiliary variables:

$$\mu_{ij} = \frac{P_i}{P_i + P_j}$$

and found P_i's such that the corresponding μ_{ij} 's are in some sense "close" to the observed r_{ij} 's. The Kullback-Leibler divergence between μ_{ij} and r_{ij} :

$$l(\vec{p}) = \sum_{i<j} n_{ij} \left[r_{ij} \log \frac{r_{ij}}{\mu_{ij}} + (1 - r_{ij}) \log \frac{1 - r_{ij}}{1 - \mu_{ij}} \right] \qquad (1)$$

is selected as the closeness measure in [10]. Minimizing this function can find P_i's. An iterative procedure is proposed to solve such constrained minimization problem:

step 1. Initialize P_i , and compute corresponding μ_{ij} ;

step 2. Repeat until convergence:

$$P_i \leftarrow P_i \cdot \frac{\sum_{j \neq i} n_{ij} r_{ij}}{\sum_{j \neq i} n_{ij} \mu_{ij}}$$

renormalize the P_i , and recomputed the μ_{ij} .

step 3. $P \leftarrow P / \sum P_i$

For simplicity, the weights are assume equal, that is, $n_{ij} = 1$ for all *i*, *j*. A simple non-iterative estimate of *P* can be obtained simply as:

$$P_i = \frac{2}{K(K-1)} \sum_{j \neq i} r_{ij}\ ,\ i, j = 1, 2, \cdots, K \qquad (2)$$

Let the posterior probabilities of $\vec{x}$ be $\vec{P}(\vec{x}) = (P_1, \cdots P_K)$. The final decision rule is: $d(\vec{x}) = \arg\max_i [P_i(\vec{x})]$

3. METHODOLOGY

In PWC, the weight matrix $\{n_{ij}\}$ in Eqn.(1) can be re-written as the follows:

$$W = \begin{bmatrix} - & W_{1,2} & W_{1,3} & \cdots & W_{1,K} \\ W_{2,1} & - & W_{23} & \cdots & W_{2,K} \\ W_{3,1} & W_{3,2} & - & \cdots & W_{3,K} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ W_{K,1} & W_{K,2} & W_{K,3} & \cdots & - \end{bmatrix}$$

where, $W_{ij} = W_{ji},\ \ i, j = 1, \cdots, K,\ \ i \neq j$. This weight matrix reflects the influence of each pairwise classifier to the final decision for a given pattern $\vec{x}$. In PWC, the weights are all assumed to be 1, which means that all binary classifiers' contributions to the final decision are the same. In fact, given a testing pattern $\vec{x} \notin c_i \cup c_j$, the pairwise probability r_{ij} is absolutely irrelevant to $\vec{x}$ because the corresponding binary classifier which is used to discriminate class c_i and c_j has not been trained with data from the true class. Consequently, using it to find $\vec{P}$ is very likely to damage the result of the calculation. Accurately, not all binary classifiers are useful and relevant to the final decision for a given pattern, some of which are meaningless, or even harmful. However we have no idea about the information because this is what we aim at determining.

3.1. Optimal PWC classifier for Each Class

To overcome the problem mentioned above, we introduce the *Optimal PWC*. Before that, we will present such a fact: Suppose the pairwise probabilities matrix is known, given a weight matrix, we can use the conventional coupling method (see the iterative procedure in Section 2) to construct a unique PWC classifier.

Definition: The *optimal weight matrix* for class c_i is:

$$W^i = \{W_{p,q}\},$$

which satisfies:

$$\begin{cases} W_{p,q} = 1, & if \quad p = i \quad or \quad q = i \\ W_{p,q} = 0, & otherwise \end{cases}\ ,\quad p, q = 1, 2, \cdots, K, \quad p \neq q$$

Using W^i, a PWC classifier can be constructed using the iterative procedure in Section 2, which is called *Optimal PWC* for c_i (in short, *O-PWC*).

For example, considering a 5-class classification problem, the *optimal weight matrix* for class c_2 and c_4 can be represented as the follows respectively:

$$\begin{bmatrix} - & 1 & 0 & 0 & 0 \\ 1 & - & 1 & 1 & 1 \\ 0 & 1 & - & 0 & 0 \\ 0 & 1 & 0 & - & 0 \\ 0 & 1 & 0 & 0 & - \end{bmatrix} \quad \begin{bmatrix} - & 0 & 0 & 1 & 0 \\ 0 & - & 0 & 1 & 0 \\ 0 & 0 & - & 1 & 0 \\ 1 & 1 & 1 & - & 1 \\ 0 & 0 & 0 & 1 & - \end{bmatrix}$$

It means that, given a pattern $\vec{x} \in c_i$, only the binary classifiers whose one of the two involved classes is c_i are considered, while other binary classifiers are all ignored. Hence, all pairwise probabilities in *O-PWC*^{i} are all relevant to class c_i.

Given a pattern $\vec{x}$, the output of *O-PWC*^{i} can be represented as a probability vector: $\vec{P}^i(\vec{x}) = (P_1^i, P_2^i, \cdots, P_K^i)$,

$i \in \{1,2,\cdots,K\}$. Each of its components represents the probability of $\vec{x}$ belonging to the respective class.

3.2. Properties of *O-PWC*

In this sub-section, we will investigate the properties of *O-PWC* in two scenarios: 1) when a given pattern is presented to different *O-PWC*s; 2) when samples from different classes are presented to the same *O-PWC*. Fig.1(upper) and Fig.1(bottom) show the two scenarios, respectively.

Given $\vec{x} \in c_i$, the output of *O-PWC*^{i}, $\vec{P}^i$, will be an "regular" probability vector where its *i*-th component P_i^i is the largest and very remarkable, while the other components will be small and similar. However, when $\vec{x}$ is presented to *O-PWC*^{j}, $j \neq i$, the output $\vec{P}^j$ does not have such properties and all of its components are "irregular".

An approximate and intuitive explanation can be given as the follows: Given an *optimal weight matrix* W^i, a new pairwise probability matrix can be constructed: $PPM^i = W^i \cdot PPM$. If $\vec{x} \in c_i$, then the values in *i*-th row of PPM^i are all bigger probabilities, while the other rows contains only one value, each of which is very small. After using Eqn.(2), the *i*-th component will be remarkable and the other components of $\vec{P}^i$ will be small and similar comparing with *i*-th component.

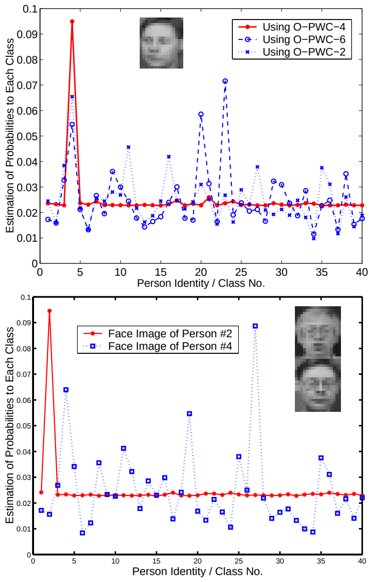


Fig. 1. Properties of *O-PWC*. The upper shows the outputs of *O-PWC*^{2}, *O-PWC*^{4} and *O-PWC*^{6} for a face image for person #4 (class c_4). The bottom shows the output of *O-PWC*^{2} for two faces from class c_2 and c_4, respectively.

For a K-class classification problem, we assume the "true" probability for each class can be represented as $\vec{T}^n$, $n \in \{1,2,\cdots,K\}$. Its component will be defined to be 1 if the label of $\vec{x}$ is *n* and 0 otherwise. For example, considering a 5-class classification problem, the "true" probabilities for c_2 can be described as: $\vec{T}^2 = (0,1,0,0,0)$. Similarly, the "true" probabilities for c_4 will be: $\vec{T}^4 = (0,0,0,1,0)$.

Based on this observations, we can see that the *O-PWC*^{i} will produce a "regular" probability vector which is closest to $\vec{T}^i$ for $\vec{x}$ if and only if $\vec{x} \in c_i$.In this paper, the following two metrics: cross entropy and square error, are used to assess the extent of "closeness":

$$CE\left(\vec{P}(\vec{x}),\vec{T}^n\right) = -\sum_{m=1}^{K} T_m^n \log_2 \frac{P_m(\vec{x})}{T_m^n} \qquad \textbf{(3)}$$

$$SE\left(\vec{P}(\vec{x}),\vec{T}^n\right) = \left\|\vec{P}(\vec{x}) - \vec{T}^n\right\| = \sum_{m=1}^{K} \left(P_m(\vec{x}) - T_m^n\right)^2 \qquad \textbf{(4)}$$

where $n \in \{1,2,\cdots,K\}$.

Based on the above analysis and observation, given a pattern $\vec{x} \in c_i$, the cross entropy or square error between $\vec{T}^i$ and $\vec{P}^i(\vec{x})$ will be minimal. So *O-PWC*^{i} can be regarded as the optimal classifier for class c_i in the sense of the cross entropy or square error.

3.3. Classification Using *O-PWC*s

For a K-class classification problem, we propose to use an array of K *O-PWC*s to perform the classification task. The systemic diagram is shown in Fig.2.

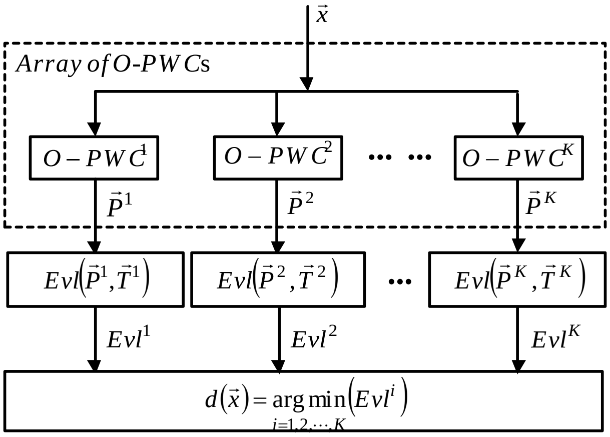

$\vec{x}$
$Array\ of\ O\text{-}PWCs$
$O-PWC^1$
$O-PWC^2$
••• •••
$O-PWC^K$
$\vec{P}^1$
$\vec{P}^2$
$\vec{P}^K$
$Evl\left(\vec{P}^1,\vec{T}^1\right)$
$Evl\left(\vec{P}^2,\vec{T}^2\right)$
•••
$Evl\left(\vec{P}^K,\vec{T}^K\right)$
Evl^1
Evl^2
Evl^K
$d(\vec{x}) = \arg\min_{i=1,2,\cdots,K}\left(Evl^i\right)$

Fig. 2. The system diagram of combination of K *O-PWC*s.

There are K channels in Fig.2, each of which will use the respective *O-PWC* to produce a probability vector for a given pattern $\vec{x}$. For a K-class problem, there will be K probability vector outputs:

$$\left\{\vec{P}^m(\vec{x})\right\},\ m = 1,2,\cdots,K$$

and at the same time the *i*-th channel is corresponding to class c_i and has a "true" probability vector $\vec{T}^i$.

For a given sample $\vec{x}$, an array of cross entropy or square error for each channel will be computed:

$$\left\{Evl^i = Evl\left(\vec{P}^i, \vec{T}^i\right)\right\},\ i = 1, 2, \cdots, K$$

where the metric function $Evl\left(\vec{P}^i, \vec{T}^i\right)$ can be square error (See Eqn.(3)) or cross entropy (See Eqn.(4)).

Based on the above analysis and observation, if $\vec{x} \in c_i$, then the output of *O-PWC*^{i} , $\vec{P}^i$, will be "regular" and be the closest to its "true" probability vector $\vec{T}^i$, the corresponding Evl^i will be minimal. However the outputs of other *O-PWC*s for $\vec{x} \in c_i$ will be "irregular". Classifying $\vec{x}$ equals to find the class label corresponding to the "regular" output.

The final decision can be represented as:

$$d(\vec{x}) = \arg\min_i \left\{Evl^i(\vec{x})\right\}$$

3.4. Performance Evaluation

Like MaxVoting and conventional PWC, our method need to evaluate K(K-1)/2 pairwise binary classifiers for a given pattern $\vec{x}$ which is time-consumed. Our method still needs to perform K coupling processes which are very fast. Hence, comparing with conventional PWC, our method does not increase computational cost too much.

3.5. Construction of pairwise probabilities matrix Using SVM

All the discussion above supposes the existence of a pairwise probabilities matrix. In fact, pairwise probability matrix can be constructed using any binary classifier which can produce probabilistic outputs. SVM is used in this paper as the pairwise binary classifier due to its better generalization ability [11]. However standard SVM can not provide a calibrated poster probability. J.Platt in [5] proposed a "SVM+Sigmoid" method to map the outputs of a binary SVM to posterior probabilities.

Given a two class classification problem, J.Platt argues that the class-conditional densities between the margins are apparently exponential and can be represented using a parametric form of a sigmoid:

$$P(y = 1 \mid f) = \frac{1}{1 + \exp(Af + B)}$$

The parameters A and B are found by minimizing the cross entropy of the training data:

$$\min -\left[\sum_i t_i \log(p_i) + (1 - t_i)\log(1 - p_i)\right]$$

where,

$$p_i = \frac{1}{1 + \exp(Af_i + B)},\ t_i = \frac{y_i + 1}{2}$$

The "SVM+Sigmoid" model leaves SVM unchanged, and is similar to Logistic Regression. In this paper, this model is adopted to map SVM outputs to posterior probabilities.

For a K-class problem, K(K-1)/2 sigmoid models will be trained. Going through all the pairwise binary SVMs, we can construct a pairwise probability matrix.

4. EXPERIMENTS

We illustrate our algorithm on the ORL face database, which consists of 400 images of 40 individuals, containing quite a high degree of variability in expression, pose and facial details. Some samples from this database are depicted in Fig.3.

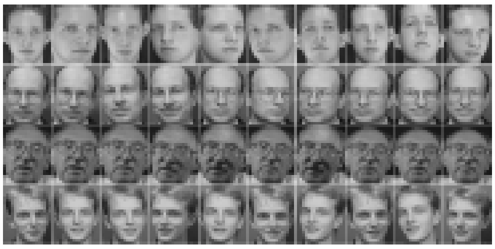


Fig. 3. Four individuals (each in one row) in the ORL. There are 10 images for each person.

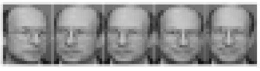
In our face recognition experiments, feature extraction phase can be performed as: we perform wavelet transform twice on the image to get the low frequency components and then whiten in order to make each vector 0-mean and 1-variance. Our extracted feature is 168 dimensions. We select 200 samples (5 for each individual) randomly as the training set. The remaining 200 samples are used as the testing set. Training phase includes two parts in this paper, one is to train 40*(40-1)/2=780 binary SVM classifiers, the other is to fit 780 sigmoid models to estimate the posterior probability of each binary SVM. In this paper, LIBSVM [1] is used to train binary SVMs. We adopt $\sigma = 0.3$ and C=10 (sigmoid kernel) to train all binary SVM classifiers.

Table 1 shows the comparison of different recognition methods on ORL database:

Table 1. Recognition accuracy rate comparation

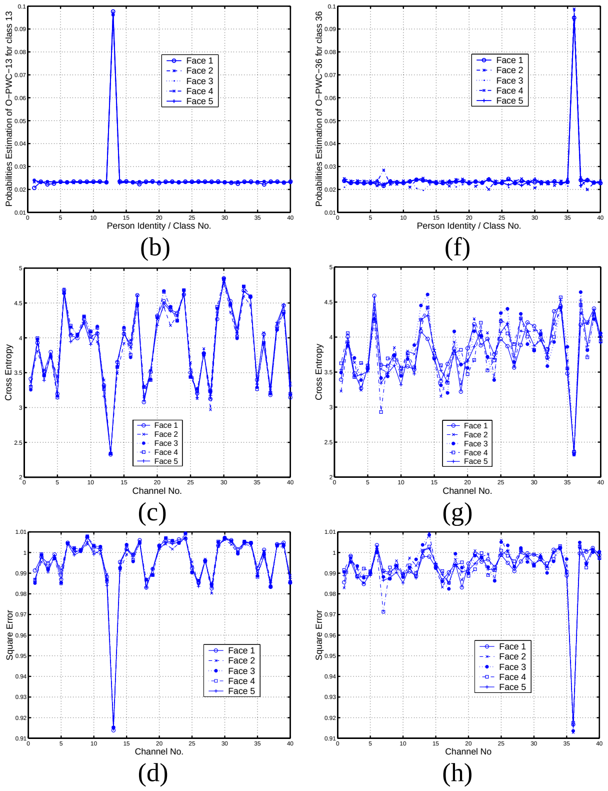
Method	MaxVoting	PWC	O-PWC (Cross Entropy)	O-PWC (Square Error)
Rate	94%	95.13%	96.79%	98.11%

It can be observed that conventional *PWC* method is superior to MaxVoting, and *O-PWC* is superior to the other two in term of accuracy rate. At the same time, we can see that, the square error metric is better than the cross entropy metric.



(a)

(e)



(b) (f)

(c) (g)

(d) (h)

Fig. 4. (a) are the five face images from person #13; (b) are the outputs of *O-PWC*^{13} for the five images in (a); (c) and (d) are the cross entropy and square error of the 40 channels for the five images in (a). Similarly, (e)-(h) are the corresponding results for the five face images shown in (b).

Fig.4 shows some results when a pattern $\vec{x}$ is presented to the framework shown in Fig.2: 1) the output of *O-PWC*^{13} for class 13, $\vec{P}^{13}$, is the closest to the "true" probabilities, $\vec{T}^{13}$; 2) The cross entropy or square error between $\vec{P}^{13}$ and $\vec{T}^{13}$, that is, the cross entropy of channel 13, will be minimal; 3) The square error metric is superior to the cross entropy metric. Hence, we can classify these five faces into class 13 correctly. Similar conclusion can be got for 5 face images from person #36. What's more, from Fig.4 we can see that, because the degree of variability in (e) is higher than that in (a), there are more variance in (f)-(h) than that in (b)-(d) for different face images from the same class.

5. CONCLUSION

This paper proposes an array of *O-PWC*s to tackle multi-class classification problems. In fact, the *optimal weight matrix* provides a method to select binary classifiers which are relevant to class c_i. *O-PWC*^{i} only considers these binary classifiers. Hence it is the optimal for c_i in the sense of cross entropy of square error. The classification accuracy rate can be improved while the computational cost will not increase too much.

6. REFERENCES

[1] Chih-Jen Lin, web site, http://csie.ntu.edu.tw/~cjlin/

[2] C.-W. Hsu and C.-J. Lin. "A comparison of methods for multi-class support vector machines", April 2001. To appear in IEEE Transactions on Neural Networks.

[3] K. Goh, E. Chang and T. Cheng, "Support Vector Machine Pairwise Classifiers with Error Reduction for Image Classification", Proceedings of ACM Intl Conf. on Multimedia (MIR Workshop), Ottawa, October 2001.

[4] M.Moreira and E.Mayoraz, "Inproving pairwise coupling classification with error correcting classifiers", Proceeding of Tenth Eruopean Conference on Machine Learning, April 1998.

[5] Platt J., "Probabilistic outputs for SVMs and comparisons to regularized likehood methods", In Advances in Large Margin Classifiers. MIT Press, 1999.

[6] Platt J, Cristianini N, Shawe-Taylor J. "Large margin DAGs for multiclass classification", In: Advances in NIPS, Vol.12, pp547~553. MIT Press, 2000.

[7] U.Krebel. "Pairwise classification and support vector machines", In "Advance in Kernel Methods – Support Vector Learning", pages 255-268, Cambridge, MA, 1999. MIT Press.

[8] Weston,J. and Watkins,C., "Multi-class support vector machines". TR CSD-TR-98-04, Dept. of Computer Science, Royal Holloway, University of London,1998.

[9] T.Dietterich and G.Bakiri. "Solving multiclass learning problems via error-correcting output codes", Journal of Artifical Intelligence Research, 2, 1995.

[10] Trevir Hastie, Robert Tibshirani, "Classification by pairwise coupling", Technical report, stanford Univ and Univ. of Toronto, 1996. in Proceeding of NIPS*97.

[11] Vapnik V., "Statistical Learning Theory", Wiley, 1998.

[12] Volker Rothm, Koji Tsuda, "Pairwise Coupling for Machine Recognition of Hand-Printed Japanese Characters", Accepted for CVPR01.