# A Supervised Feature Extraction Algorithm for Multi-class

Shifei Ding[1,2], Fengxiang Jin[3], Xiaofeng Lei[1], and Zhongzhi Shi[2]

[1] School of Computer Science and Technology, China University of Mining and Technology, Xuzhou 221008 China
[2] Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100080 China
[3] College of Geoinformation Science and Engineering, Shandong University of Science and Technology, Qingdao 266510 P.R. China
dingsf@cumt.edu.cn

**Abstract.** In this paper, a novel supervised information feature extraction algorithm is set up. Firstly, according to the information theories, we carried out analysis for the concept and its properties of the cross entropy, then put forward a kind of lately concept of symmetry cross entropy (SCE), and point out that the SCE is a kind of distance measure, which can be used to measure the difference of two random variables. Secondly, Based on the SCE, the average symmetry cross entropy (ASCE) is set up, and it can be used to measure the difference degree of a multi-class problem. Regarding the ASCE separability criterion of the multi-class for information feature extraction, a novel algorithm for information feature extraction is constructed. At last, the experimental results demonstrate that the algorithm here is valid and reliable, and provides a new research approach for feature extraction, data mining and pattern recognition.

## 1 Introduction

Feature extraction is one of the most importmant steps in pattern recognition, data mining, machine learning and so on[1,2]. In order to choose a subset of the original features by reducing irrelevant and redundant, many feature selection algorithms have been studied. The literature contains several studies on feature selection for unsupervised learning in which he objective is to search for a subset of features that best uncovers "natural" groupings (clusters) from data according to some criterion. For example, principal components analysis (PCA) is an unsupervised feature extraction method that has been successfully applied in the area of face recognition, feature extraction and feature analysis[3-5]. But the method of PCA is effective to deal with the small size and low-dimensional problems, and gets the extensive application in Eigenface and feature extraction. In high-dimensional cases, it is very difficult to compute the principal components directly[6]. Fortunately, the algorithm of Eigenfaces artfully avoids this difficulty by virtue of the singular decomposition technique. Thus, the problem of calculating the eigenvectors of the total covariance matrix, a high-dimensional matrix, is transformed into a problem of calculating the eigenvectors of a much lower dimensional matrix[7].

F.P. Preparata, X. Wu, and J. Yin (Eds.): FAW 2008, LNCS 5059, pp. 323–329, 2008.

Now an important question is how to deal with supervised information feature extraction. For supervised feature extraction problem, some authors have studied by discriminate analysis, bayes decision theory et al. But these methods depend on probability distributions of some classifications. In this paper, the authors have studied this field on the basis of these aspects. Firstly, we study and discuss the information theory, cross entropy theory, and point out its shortage. Secondly, a new concept of symmetry cross entropy (SCE) is put forward, and proved that the SCE is a kind of distance measure. At the same time, based on the SCE, we give the average SCE, i.e. ASCE. Which is regarded multi-class separability criterion. Thirdly, according to ASCE, a new information feature extraction algorithm is constructed. At last, the proposed algorithm here is tested in practice, and the experimental results indicate that it is efficient and reliable.

## 2 Feature Extraction Algorithm

In order to set up information feature extraction algorithm, we firstly discuss the following new concept of symmetry cross entropy and feature extraction theorem.

### 2.1 Symmetry Cross Entropy

Shannon[8] put forward the concept of information entropy for the very first time in 1948. The cross entropy (CE), or the relative entropy, is used for measuring difference information between the two probability distributions. But the CE satisfies only nonnegativity, normalization and dissatisfies symmetry and triangle inequation. For this reason, we carry out the improvement, and give the following definition.

**Definition 1.** Let $X$ be a discrete random variable with two probability distribution vectors $P$ and $Q$, where $P=(p_1,p_2,\cdots,p_n)$, $Q=(q_1,q_2,\cdots,q_n)$, the CE between $P$ and $Q$ is defined as

$$H(P\|Q)=\sum_{i=1}^{n}p_i\log\frac{p_i}{q_i}=E\left(\log\frac{p_i}{q_i}\right) \tag{1}$$

In the above definition denoted by formula (1), we show that the CE is always nonnegative and is zero if and only if $p_i=q_i$. However, it is not a true distance between distributions since it is not symmetric and does not satisfy the triangle inequality. In order to make it true distance between distributions, we improve the CE as follows.

**Definition 2.** Suppose that the $H(P\|Q)$ and $H(Q\|P)$ are CEs of $P$ to $Q$ and $Q$ to $P$ respectively, the symmetric cross entropy (SCE) between $P$ and $Q$, denoted by $D(P,Q)$, defined as

$$D(P,Q)=H(P\|Q)+H(Q\|P)$$

$$=\sum_{i=1}^{n}p_i\log p_i+\sum_{i=1}^{n}q_i\log q_i-\sum_{i=1}^{n}p_i\log q_i-\sum_{i=1}^{n}q_i\log p_i \tag{2}$$

It is called Symmetric Cross Entropy (SCE) of $\boldsymbol{P}$ and $\boldsymbol{Q}$.

According to the definition of the SCE, we have the following theorem.

**Theorem 1.** Suppose that the SCE is defined by formula (2), then the SCE is a kind of distance measure, i.e. $D(P,Q)$ satisfies basic properties as follows.

(Ⅰ) Non-negativity: $D(P,Q) \geq 0$, and $D(P,Q) = 0 \Leftrightarrow P = Q$;

(Ⅱ) Symmetry: $D(P,Q) = D(Q,P)$;

(Ⅲ) Triangle inequation: Suppose that $W = (w_1, w_2, \cdots, w_n)$ is another probability distribution vector of the discrete random variable $X$, then

$$D(P,Q) \leq D(P,W) + D(W,Q) \tag{3}$$

Therefore, the SCE is a distance measure, which can be used to measure the degree of variation between two random variables. The SCE is considered as separability criterion of the two-class for information feature extraction. It can be seen that the smaller the SCE is, the smaller the difference of two-class. In particular, when the SCE=0, the two-class are same completely. For information feature extraction, under the condition of the given reduction dimensionality denoted by $d$, we should select such $d$ characteristics that make the value of the SCE approach maximum. For convenience, we use the following function, denoted by $H(P,Q)$, in instead of above the SCE.

$$H(P,Q) = \sum_{i=1}^{n} (p_i - q_i)^2 \tag{7}$$

For a multi-class problem, based on the formula (4), the SCE is computed for every class $i$ and $j$, where $i$ and $j$ denote number of class

$$H_{ij} = \sum_{k=1}^{n} (p_k^{(i)} - p_k^{(j)})^2 \tag{5}$$

The average symmetric cross entropy (ASCE) can be expressed as follows

$$H = \sum_{i=1}^{M} \sum_{j=1}^{M} p_k^{(i)} p_k^{(j)} d_{ij} = \sum_{i=1}^{M} \sum_{j=1}^{M} \sum_{k=1}^{n} p_k^{(i)} p_k^{(j)} (p_k^{(i)} - p_k^{(j)})^2 \tag{6}$$

being equivalent to the SCE, we should select such $d$ characteristics that make the value of $H$ approach maximum. In fact, $H$ approaching maximum is equivalent to $H_{ij}$ approaching maximum, so information feature extraction for a multi-class problem is also equivalent to a two-class problem.

In order to set up the information feature extraction algorithm, we first give the following theorem.

**Theorem 2.** Suppose $\{x_j^{(1)}\}$ $(j=1,2,\cdots,N_1)$ and $\{x_j^{(2)}\}$ $(j=1,2,\cdots,N_2)$ with covariance matrices $G^{(1)}$ and $G^{(2)}$ are squared normalization feature vectors, so-called squared normalization indicates

$$\sum_{k=1}^{n}(x_{jk}^{(i)})^2=1 \tag{7}$$

where $x_{jk}^{(i)}(i=1,2)$ denotes the kth feature component of the feature vector $x_j^{(i)}$. Then the SCE, i.e. the $H(P,Q)$ =maximum if and only if the coordinate system is composed of $d$ eigenvectors corresponding to the first $d$ eigenvalues of the matrix $A=G^{(1)}-G^{(2)}$.

So for $\{\boldsymbol{X}_j^{(1)}\}$ $(j=1,2,\cdots,N_1)$ and $\{X_j^{(2)}\}$ $(j=1,2,\cdots,N_2)$ $(j=1,2,\cdots,N_2)$ with covariance matrices $G^{(1)}$ and $G^{(2)}$ are squared normalization feature vectors. The k-th feature component of $\boldsymbol{X}_{\boldsymbol{j}}^{(\boldsymbol{i})}$ is denoted by $x_{jk}^{(i)}(i=1,2;k=1,2,\cdots,n)$, and the square mean of each component for every class is $\gamma_k^{(i)}=\frac{1}{N_i}\sum_{j=1}^{N_i}(x_{jk}^{(i)})^2$, where $i=1,2;j=1,2,\cdots,n$. Obviously $\gamma_k^{(i)}\geq 0$, and then

$$\sum_{k=1}^{n}\gamma_k^{(i)}=\sum_{k=1}^{n}\frac{1}{N_i}\sum_{j=1}^{N_i}(x_{jk}^{(i)})^2=\sum_{j=1}^{N_i}\frac{1}{N_i}\sum_{k=1}^{n}(x_{jk}^{(i)})^2=\sum_{j=1}^{N_i}\frac{1}{N_i}=1 \tag{8}$$

Namely $\gamma_k^{(i)}\geq 0$ and $\sum_{k=1}^{n}\gamma_k^{(i)}=1$. Therefore, we can comprehend $\{\gamma_k^{(i)}\}$ as the probability distribution defined by $\boldsymbol{X}_{\boldsymbol{j}}^{(\boldsymbol{i})}$. Suppose that the $(k,l)$ element of symmetric matrix $\boldsymbol{G}^{(\boldsymbol{i})}$ $(i=1,2)$ is $g_{kl}^{(i)}=\frac{1}{N_i}\sum_{j=1}^{N_i}x_{jk}^{(i)}x_{jl}^{(i)}$. Record $\gamma^{(i)}=(\gamma_1^{(i)},\gamma_2^{(i)},\cdots,\gamma_n^{(i)})$, then every components of $\gamma^{(i)}$ is element of $\boldsymbol{G}^{(\boldsymbol{i})}$ $(i=1,2)$ in diagonal line. Let

$$s=s(\gamma^{(1)},\gamma^{(2)})=\sum_{k=1}^{n}(\gamma_k^{(1)}-\gamma_k^{(2)})^2 \tag{9}$$

### 2.2 Feature Extraction Algorithm

Suppose three classes $C_1$, $C_2$, and $C_3$ with covariance matrices $G^{(1)}, G^{(2)}$ and $G^{(3)}$ are squared normalization feature vectors. According to the discussion above, an

algorithm of information feature extraction based on the ASCE is derived and is as follows.

Step 1. Data pretreatment. Perform square normalization transformation for two classes original data according to the formula (7), and get data matrix $x^{(1)}, x^{(2)}, x^{(3)}$ respectively.

Step 2. Compute symmetric matrix $A, B, C$. Calculate the covariance matrixes $G^{(1)}, G^{(2)}, G^{(3)}$ and then get symmetric matrix as follows.

$$A = G^{(1)} - G^{(2)}\ ,\ B = G^{(1)} - G^{(3)}\ ,\ C = G^{(2)} - G^{(3)} \tag{10}$$

Step 3. Calculate all eigenvalues and corresponding eigenvectors of the matrix $A$ according to Jacobi method.

Step 4. Construct extraction index. The total sum of variance square is denoted by

$$V_n = \sum_{k=1}^{n} \lambda_k^2\ ,\ V_d = \sum_{k=1}^{d} \lambda_k^2 \tag{11}$$

and then the variance square ratio (VSR) is VSR= $V_d / V_n = V_d / s_0$. The VSR value can be used to measure the degree of information extraction. Generally speaking, so long as $V_i \geq 80\%$, the purpose of feature extraction is reached.

Step 5. Construct extraction matrix. When $V_i \geq 80\%$, we select $d$ eigenvectors corresponding to the first $d$ eigenvalues, and construct the information extraction matrix $T = (u_1, u_2, \cdots, u_d)$.

Step 6. Feature extraction. The data matrixes $x^{(1)}, x^{(2)}, x^{(3)}$ is transformed by

$$y^{(i)} = T' x^{(i)} (i = 1,2,3) \tag{12}$$

and the purpose to compress the data information is attained.

### 2.3 Experimental Results

The original data sets come from reference[9], they are divided into three classes $C_1$, $C_2$, and $C_3$, and denote light occurrence, middle occurrence, and heavy occurrence about the occurrence degree of the pests respectively.

According to the algorithm set up above, and applying the DPS data processing system, the compressed results for three classes are expressed in Fig. 1.

Fig.1. shows that the distribution of feature vectors after compressed for the class $C_1$ denoted by "+", the class $C_2$ denoted "*" and the class $C_3$ denoted "^", is obviously concentrated relatively, meanwhile for these three classes, the within-class distance is small, the between-class distance is big, and the ASCE is maximum. Therefore, 2-dimensional pattern vector loaded above 99% information contents of the original 5-dimensional pattern vector. The experimental results demonstrate that the algorithm presented here is valid and reliable, and takes full advantage of the class-label information of the training samples.

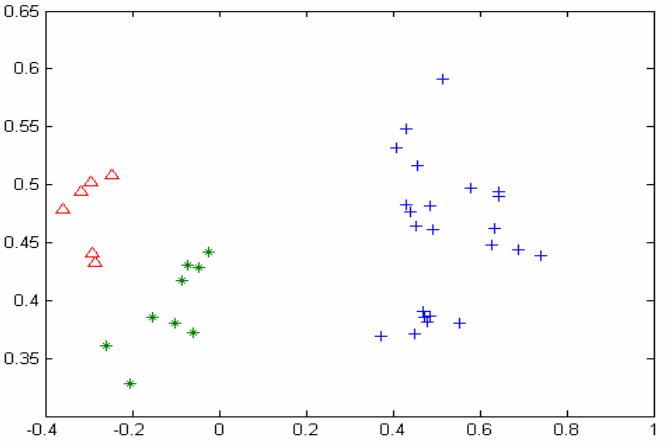


**Fig. 1.** The compressed results for three classes

## 3 Conclusions

From the information theory, studied and discussed the compression problem of the information feature in this paper, and come to a conclusion. According to the definition of the CE, a new concept of the SCE is proposed, and proved that the SCE is a distance measure which can be used to measure the degree of two-class random variables. The average SCE (ASCE) is given based on SCE, and it is to measure the difference degree for the multi-class problem. Regarding the ASCE separability criterion of the multi-class for information feature compression, we design a novel information feature compression algorithm. The experimental results show that algorithm presented here is valid, and compression effect is significant.

## Acknowledgements

This work is supported by the National Natural Science Foundation of China under Grant No.40574001, 863 National High-Tech Program under Grant No. 2006AA01Z128, and the Opening Foundation of the Key Laboratory of Intelligent Information Processing of Chinese Academy of Sciences, under Grant No.IIP2006-2.

## References

1. Duda, R.O., Hart, P.E. (eds.): Pattern Classification and Scene Analysis. Wiley, New York (1973)
2. Fukunaga, K. (ed.): Introduction to Statistical Pattern Recognition, 2nd edn. Academic Press, London (1990)
3. Ding, S.F., Shi, Z.Z.: Supervised Feature Extraction Algorithm Based on Improved Polynomial Entropy. Journal of Information Science 32(4), 309–315 (2006)
4. Hand, D.J. (ed.): Discrimination and Classification. Wiley, New York (1981)

5. Nadler, M., Smith, E.P. (eds.): Pattern Recognition Engineering. Wiley, New York (1993)
6. Yang, J., Yang, J.Y.: A Generalized K-L Expansion Method That Can Deal With Small Sample Size and High-dimensional Problems. Pattern Analysis Applications 6(6), 47–54 (2003)
7. Zeng, H.L., Yu, J.B., Zeng, Q.: System Feature Reduction on Principal Component Analysis. Journal of Sichuan Institute of Light Industry and Chemical Technology 12(1), 1–4 (1999)
8. Shannon, C.E.: A Mathematical Theory of Communication. Bell Syst. Tech. J. 27, 379–423 (1948)
9. Tang, Q.Y., Feng, M.G. (eds.): Practical Statistics and DPS Data Processing System. Science Press, Beijing (2002)