A Novel Supervised Information Feature Compression Algorithm

Shifei Ding^{1,2} and Zhongzhi Shi²

¹ College of Information Science and Engineering,
Shandong Agricultural University, Taian 271018 P.R. China

² Key Laboratory of Intelligent Information Processing, Institute of Computing Technology,
Chinese Academy of Sciences, Beijing 100080 P.R.China
sfding@sdau.edu.cn; shizz@ics.ict.ac.cn

Abstract. In this paper, a novel supervised information feature compression 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, We make the SCE separability criterion of the classes for information feature compression, and design a novel algorithm for information feature compression. At last, the experimental results demonstrate that the algorithm here is valid and reliable, and provides a new research approach for feature compression, data mining and pattern recognition.

1 Introduction

Feature extraction or compression is one of the most important steps in pattern recognition, data mining, machine learning and so on. In order to choose a subset of the original features by reducing irrelevant and redundant, many feature selection algorithms have been studied[1,2]. The literature contains several studies on feature selection for unsupervised learning in which the 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[3,4]. But the method of PCA is effective to deal with the small size and high-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[5].

Now an important question is how to deal with supervised information feature compression. In this paper, the authors are going on studying this field on the basis of these studies. Firstly, we study and discuss the cross entropy theory, and point out its shortage, then put forward a new concept of symmetry cross entropy (SCE), and prove that the SCE is a kind of distance measure. Secondly, we regard SCE as class separability criterion, and design a new algorithm of information

feature compression. At last, our experimental results indicate that the proposed algorithm here is efficient and reliable.

2 Feature Compression Algorithm

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

2.1 Symmetry Cross Entropy

Shannon put forward the concept of information entropy for the very first time in 1948[6]. 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 Suppose that $\mathbf{P} = (p_1, p_2, \dots, p_n)$ and $\mathbf{Q} = (q_1, q_2, \dots, q_n)$ are two probability vectors of discrete random variable $\mathbf{X}$, and $H(\mathbf{P} \parallel \mathbf{Q})$ and $H(\mathbf{Q} \parallel \mathbf{P})$ are CE of $\mathbf{P}$ to $\mathbf{Q}$ and $\mathbf{Q}$ to $\mathbf{P}$ respectively. Then we define

$$D(\mathbf{P}, \mathbf{Q}) = H(\mathbf{P} \parallel \mathbf{Q}) + H(\mathbf{Q} \parallel \mathbf{P}) \quad (1)$$

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

We can prove that the SCE satisfies three basic properties as follows.

- 1) Nonnegativity: $D(\mathbf{P}, \mathbf{Q}) \geq 0$, $D(\mathbf{P}, \mathbf{Q}) = 0 \Leftrightarrow \mathbf{P} = \mathbf{Q}$;
- 2) Symmetry: $D(\mathbf{P}, \mathbf{Q}) = D(\mathbf{Q}, \mathbf{P})$;
- 3) Triangle inequation: Suppose that $\mathbf{W} = (w_1, w_2, \dots, w_n)$ is another probability vector, then $D(\mathbf{P}, \mathbf{Q}) \leq D(\mathbf{P}, \mathbf{W}) + D(\mathbf{W}, \mathbf{Q})$.

Therefore, the SCE is a kind of distance measurement, which can be used to measure the difference of two random variables. In order to compute conveniently, we can use the function as follows in accordance with the above $D(\mathbf{P}, \mathbf{Q})$, i.e.

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

Doing like this, we don't affect the results to select d optimal features.

2.2 Compression Theorem

Suppose that $\{\mathbf{X}_j^{(1)}\}$ ($j=1,2,\dots,N_1$) and $\{\mathbf{X}_j^{(2)}\}$ ($j=1,2,\dots,N_2$) are squared normalization pattern vectors which belongs to two classes. The k th feature component of $\mathbf{X}_j^{(i)}$ is denoted by $x_{jk}^{(i)}$ ($i=1,2; k=1,2,\dots,n$). 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 $i = 1, 2; j = 1, 2, \dots, n$.

Obviously $\gamma_k^{(i)} \geq 0$, and then

$$\sum_{k=1}^n \gamma_k^{(i)} = \sum_{k=1}^n \frac{1}{N_i} \sum_{j=1}^{N_i} (x_{jk}^{(i)})^2 = \sum_{j=1}^{N_i} \frac{1}{N_i} \sum_{k=1}^n (x_{jk}^{(i)})^2 = \sum_{j=1}^{N_i} \frac{1}{N_i} = 1 \quad (3)$$

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 $X_j^{(i)}$. Suppose that the (k, l) element of symmetric matrix $G^{(i)}$ ($i = 1, 2$) is $g_{kl}^{(i)} = \frac{1}{N_i} \sum_{j=1}^{N_i} x_{jk}^{(i)} x_{jl}^{(i)}$. Record $\gamma^{(i)} = (\gamma_1^{(i)}, \gamma_2^{(i)}, \dots, \gamma_n^{(i)})$, then every components of $\gamma^{(i)}$ is element of $G^{(i)}$ ($i = 1, 2$) in diagonal line. Let

$$s = s(\gamma^{(1)}, \gamma^{(2)}) = \sum_{k=1}^n (\gamma_k^{(1)} - \gamma_k^{(2)})^2 \quad (4)$$

According to discussion above, we can get the following compression theorem.

Theorem 1 Suppose that $A = G^{(1)} - G^{(2)}$, and the function s defined by formula (4), then the SCE=maximum if and only if the coordinate system as s =maximum is composed of the d eigenvectors in correspondence with eigenvalues satisfied definite conditions of the matrix A .

2.3 Feature Compression Algorithm

In order to explain the effect of information feature compression, we make the eigenvalues λ_i ($i = 1, 2, \dots, n$) of matrix A arrange $\lambda_1^2 \geq \lambda_2^2 \geq \dots \geq \lambda_d^2 \geq \dots \geq \lambda_n^2$. The

total variance sum of square is denoted by $V_n = \sum_{k=1}^n \lambda_k^2$, and then the variance square

ratio is defined as $V = V_d / V_n$, which can measure the degree of information compression. According to the discussion above, we can get the algorithm of information feature compression based on SCE as follows.

Step 1 Carry out square normalization processing for the two classes original data, and get data matrix $x^{(i)}$ ($i = 1, 2$).

Step 2 Carry out centralization for obtained data matrix x , and then calculate the symmetric matrix $G^{(i)}$ ($i = 1, 2$) and difference matrix A .

Step 3 Calculate all eigenvalues corresponding to all eigenvectors of the matrix $A = G^{(1)} - G^{(2)}$.

Step 4 Construct information compression matrix $T = (u_1, u_2, \dots, u_d)$.

Step 5 Compress data matrix x based on $y^{(i)} = T'x^{(i)} (i = 1, 2)$, and so we reach the purpose of optimal compression of information feature.

3 Conclusions

According to the definition of cross entropy, we propose a new concept of the SCE, and point out that the SEC is a kind of distance measure. Based on SCE, we set up a novel separability criterion, which can be used to measure the difference degree between the random variables, and construct a compression algorithm for information feature based on SCE.

Acknowledgements

This study is supported by the National Natural Science Foundation of China under grant No.60435010, the China Postdoctor Science Foundation under grant No.2005037439, and the Doctoral Science Foundation of Shandong Agricultural University under grant No. 23241.

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. Academic Press, 2nd ed., New York (1990)
3. Hand, D.J. (ed.): Discrimination and Classification. Wiley, New York (1981)
4. Nadler, M., Smith, E.P. (eds.): Pattern Recognition Engineering. Wiley, New York (1993)
5. 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) (2003) 47-54
6. Shannon, C.E.: A Mathematical Theory of Communication. Bell Syst. Tech. J. 27 (1948) 379-423