Supervised feature extraction algorithm based on improved polynomial entropy

Shifei Ding
College of Information Science and Engineering, Shandong Agricultural University, Taian, 271018, China
Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100080, China

Zhongzhi Shi
Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100080, China

Received 2 September 2005
Revised 25 October 2005

Abstract.
Based on information entropy theory, a novel feature extraction algorithm based on improved polynomial entropy (IPE) is set up. Firstly, the concepts and their properties of information entropy and cross entropy (CE) are analysed and studied. On this foundation, we prove that symmetrical cross entropy (SCE) proposed here based on CE satisfies three axioms of the distance, i.e. nonnegativity, symmetry and triangle inequation. So SCE is a kind of distance measure, which can be used to measure the degree of variation between two random variables. Secondly, for convenience, we propose a new concept of improved polynomial entropy (IPE) based on polynomial entropy (PE), and explain that IPE is equivalent to SCE. Thirdly, we make IPE separability criterion of the classes for feature extraction, and design a novel feature extraction algorithm based on IPE. Finally, the experimental results demonstrate that the algorithm proposed here is valid and reliable, and it provides a new research approach for feature extraction, data mining and pattern recognition.

Keywords: information entropy theory; cross entropy (CE); symmetrical cross entropy (SCE); polynomial entropy (PE); improved polynomial entropy (IPE); supervised feature extraction

1. Introduction

It is well known that the discovery and selection of an effective feature set are most important and difficult problems in pattern recognition (PR). The choice of features to represent the patterns affects several aspects of the PR algorithm, such as accuracy, required learning time, and the necessary numbers of samples [1, 2]. The difficulties in selecting effective features are due to several facts. Firstly, it is necessary to select a feature combination that separates the class volumes so that the classes can be distinguished, but this property of a feature set is not generally obvious without carrying out a PR or data visualization analysis of the training set using the selected features. Secondly, there are often many features, and combinatorial large numbers of feature combinations to select from them. Note that the number of feature subset combinations with m features from a collection of N total features is N!/[(m![(N – m))]. One might expect that the inclusion of increasing numbers of features would increase the likelihood of including enough information to separate the class
volumes. Unfortunately, this is not true if the size of the training data set does not also increase rapidly with each additional feature included. This is the so-called ‘curse of dimensionality’ [3–5]. In order to choose a subset of the original features by reducing irrelevant and redundant items, many feature selection algorithms have been studied.

The literature contains several studies on feature selection for unsupervised learning in which the main objective is to search for a subset of features that best uncovers ‘natural’ groupings (clusters) from data according to some criterion [6]. 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 [7, 8]. But the method of PCA is effective to deal with small-size and high-dimensional problems, and gets extensive application in eigenface and feature extraction [9]. In high-dimensional cases, it is very difficult to compute the principal components directly. 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.

Now an important question is how to deal with supervised feature extraction? Some researchers have studied this [10, 11]. In this paper, the authors are going to study this field on the basis of these studies. We study and discuss information entropy and cross entropy (CE), and point out their shortfalls. In order to overcome these shortfalls, we set up symmetrical cross entropy (SCE) and prove that SCE is a kind of distance measure. In addition, a new concept of improved polynomial entropy (IPE) based on polynomial entropy (PE) is proposed, which is equivalent to SCE. We regard IPE as a criterion of class separability, and design a new algorithm of supervised feature extraction.

2. Supervised feature extraction algorithm based on IPE

2.1. Definition of entropy

We first introduce the concept of entropy proposed by Shannon [12] in 1948, which is a measure of uncertainty of a random variable. Let $X$ be a discrete random variable with two probability distribution vectors $P$ and $Q$, where

$$P = (p_1, p_2, \ldots, p_n), \quad p_i \geq 0 \quad (i = 1, 2, \ldots, n),$$

and

$$\sum_{i=1}^{n} p_i = 1$$

(1)

$$Q = (q_1, q_2, \ldots, q_n), \quad q_i \geq 0 \quad (i = 1, 2, \ldots, n),$$

and

$$\sum_{i=1}^{n} q_i = 1$$

(2)

**Definition 1.** Suppose that $X$ is a discrete random variable with probability distribution vector $P$ denoted by formula (1), then the entropy, or Shannon entropy $H(X)$ of a discrete random variable $X$ is defined by

$$H(X) = -\sum_{i=1}^{n} p_i \log p_i$$

(3)

We also write $H(p)$ for the above quantity. The log is to the base 2 and entropy is expressed in bits. We will use the convention that $0 \log 0 = 0$. If the base of the logarithm is $e$, then the entropy is measured in nats and if the base of the logarithm is 10, then the entropy is measured in harts. It turns out that $H(X)$ can be thought of as a measure of the following things about $X$:

(a) the amount of ‘information’ provided by an observation of $X$;
(b) our ‘uncertainty’ about $X$;
(c) the ‘randomness’ of $X$.

Shannon had proved that $0 \leq H(X) \leq \log n$, and $H(X) = 0$ if and only if (iff) $p_i = 1$ for some $i$, $H(X) = \log n$ iff $p_i = 1/n$ for all $i$.

2.2. Cross entropy

Cross entropy (CE), or relative entropy, or Kullback-Leibler distance is a measure of the distance between two probability distributions of random variable. In statistics, it arises as an expected logarithm of the likelihood ratio [13, 14]. The CE is a measure of the inefficiency of assuming that the distribution is $Q$ when the true distribution is $P$.

**Definition 2.** Suppose that $X$ is a discrete random variable with two probability distribution vectors $P$ and $Q$, then the cross entropy between $P$ and $Q$ denoted by $D(P\|Q)$ is defined as

$$D(P\|Q) = \sum_{i=1}^{n} p_i \log \frac{p_i}{q_i} = \sum_{i=1}^{n} p_i \log p_i - \sum_{i=1}^{n} p_i \log q_i$$

$$= -f(p, p) + f(p, q) = -H(p) + f(p, q)$$

(4)

where
In the above definition we use the convention that 
\( 0 \log(0/q) = 0 \) and \( p \log(p/0) = \infty \). But cross entropy satisfies only non-negativity, and dissatisfies symmetry and triangle inequation, and so the CE should strictly be a generalized distance. For this reason, we improve CE, and give the definition of symmetrical cross entropy (SCE) as follows.

**Definition 3.** Suppose that \( X \) is a discrete random variable with two probability distribution vectors \( P \) and \( Q \), then the SCE denoted by \( D(P \Vert Q) \) is defined as

\[
D(P, Q) = H(P \Vert Q) + H(Q \Vert P)
\]

\[
= \sum_{i=1}^{n} p_i \log p_i + \sum_{i=1}^{n} q_i \log q_i - \sum_{i=1}^{n} p_i \log q_i - \sum_{i=1}^{n} q_i \log p_i
\]

(5)

where

\[
f(p, q) = \sum_{i=1}^{n} p_i \log q_i, f(q, p) = H(Q) = -\sum_{i=1}^{n} q_i \log q_i.
\]

In order to prove the SCE is a kind of distance measure, we first give the following Lemma.

**Lemma 1** (15). Let \( X \) be a discrete random variable with two probability distribution vectors \( P \) and \( Q \). If and only if \( P = Q \), then

\[
H(X) = -\sum_{i=1}^{n} p_i \log p_i \leq -\sum_{i=1}^{n} q_i \log q_i
\]

(6)

**Theorem 1.** Suppose that the SCE, \( D(P, Q) \), is defined by formula (5), and then \( D(P, Q) \) is a kind of distance measure, namely, it satisfies the following three axioms.

**Axiom 1.** Non-negativity: \( D(P, Q) \geq 0 \) and \( D(P, Q) = 0 \) if and only if \( P = Q \).

**Axiom 2.** Symmetry: \( D(P, Q) = D(Q, P) \).

**Axiom 3.** Triangle inequation: suppose that \( W = (w_1, w_2, \ldots, w_n) \) is another probability vector of the discrete random variable \( X \), where \( w_i \geq 0 (i = 1, 2, \ldots, n) \) and \( \sum_{i=1}^{n} w_i = 1 \), then \( D(P, Q) \leq D(P, W) + D(W, Q) \).

**Proof.** Axiom 1 and Axiom 2 are obviously right and Axiom 3 is proved as follows. Based on the definition of \( D(P, Q) \), we have

\[
D(P, Q) = \sum_{i=1}^{n} p_i \log p_i + \sum_{i=1}^{n} q_i \log q_i -
\]

(7)

\[
\sum_{i=1}^{n} p_i \log q_i - \sum_{i=1}^{n} q_i \log p_i
\]

\[
D(P, W) = \sum_{i=1}^{n} p_i \log p_i + \sum_{i=1}^{n} w_i \log w_i -
\]

(8)

\[
\sum_{i=1}^{n} p_i \log w_i - \sum_{i=1}^{n} q_i \log p_i
\]

\[
D(W, Q) = \sum_{i=1}^{n} w_i \log w_i + \sum_{i=1}^{n} q_i \log q_i -
\]

(9)

\[
\sum_{i=1}^{n} w_i \log q_i - \sum_{i=1}^{n} q_i \log w_i
\]

According to formulae (7), (8) and (9), we can get

\[
D(P, W) + D(W, Q) - D(P, Q) = 2 \sum_{i=1}^{n} w_i \log w_i
\]

(10)

\[
- \sum_{i=1}^{n} w_i \log q_i - \sum_{i=1}^{n} q_i \log w_i - \sum_{i=1}^{n} p_i \log w_i
\]

According to Lemma 1, i.e. formula (6), we get

\[
\sum_{i=1}^{n} w_i \log w_i \geq \sum_{i=1}^{n} q_i \log q_i
\]

(11)

\[
\sum_{i=1}^{n} w_i \log w_i \geq \sum_{i=1}^{n} q_i \log w_i
\]

(12)

Putting formulae (11) and (12) into formula (10), we get

\[
D(P, W) + D(W, Q) - D(P, Q) \geq - \sum_{i=1}^{n} q_i \log w_i
\]

(13)

\[
- \sum_{i=1}^{n} p_i \log w_i + \sum_{i=1}^{n} p_i \log q_i + \sum_{i=1}^{n} q_i \log p_i
\]

\[
= \sum_{i=1}^{n} p_i \log w_i + \sum_{i=1}^{n} q_i \log w_i
\]

In 1968, P. Nath gave three measures of probability vector [16], namely

\[
H(P, Q, W) = \sum_{i=1}^{n} p_i \log \frac{q_i}{w_i} \geq 0
\]

(14)

Putting formula (14) into formula (13), we have

\[
D(P, W) + D(W, Q) - D(P, Q) \geq 0
\]

(15)

According to formula (15), we see that Axiom 3 is finished.
2.3. Improved polynomial entropy

Watanabe [17], in 1983, gave another definition of entropy. He used polynomial function \( g(x) = -x^2 \) instead of logarithm function \( f(x) = -x \log x \), and considered the following polynomial function

\[
P(X) = P(p) = 1 - \sum_{i=1}^{n} p_i^2
\]  

(16)

In order to distinguish it from Shannon entropy, PE is very simple and easy to use in practice. Being similarly to Shannon entropy, PE also satisfies 0 ≤ P(X) ≤ (n-1)/n. Furthermore P(X) = 0 iff \( p_i = 1 \) for some \( i \), and P(X) = (n-1)/n iff \( p_i = 1/n \) for all \( i \).

According to PE, an improved PE (IPE) can be obtained:

\[
H(P, Q) = g(p, q) + g(q, p) - g(p, p) - g(q, q)
\]

\[
= g(p, q) + g(q, p) - P(p) - P(q) = \sum_{i=1}^{n} (p_i - q_i)^2
\]  

(17)

where

\[
g(p, q) = 1 - \sum_{i=1}^{n} p_i q_i, g(p, p) = P(p) = 1 - \sum_{i=1}^{n} p_i^2
\]

and

\[
g(q, q) = P(q) = 1 - \sum_{i=1}^{n} q_i^2
\]

(18)

Being equivalent to D(P, Q), IPE, that is H(P, Q), also satisfies non-negativity, symmetry and triangle inequality. Therefore, the IPE is a kind of distance measure too, which can be used to measure the degree of variation between two random variables. We think of the IPE as a separability criterion of the classes for feature extraction. We can see that the smaller IPE, the smaller the difference between two groups of data. In particular, when the value of IPE is zero, the two groups of data are the same completely, in other words there is no difference at this time. For feature extraction, under the condition of the given reduction dimensionality denoted by \( d \), we should select \( d \) features, and make the value of IPE reach the maximum.

2.4. Supervised feature extraction algorithm based on IPE

2.4.1. Basic algorithm principle. Suppose that \( \{X_i^{(1)}\} (j = 1, 2, \ldots, N_1) \) and \( \{X_i^{(2)}\} (j = 1, 2, \ldots, N_2) \) are squared normalization pattern vectors which belong to two classes, while the so-called squared normalization pattern vector means

\[
\sum_{k=1}^{N} (x_{jk}^{(i)})^2 = 1
\]  

(19)

The \( k \)th feature component of \( X_i^{(i)} \) is denoted by \( x_{jk}^{(i)}(i = 1, 2; k = 1, 2, \ldots, n) \). The squared mean of each component for every class is

\[
\gamma_k^{(i)} = 1 - \frac{1}{N_i} \sum_{j=1}^{N_i} (x_{jk}^{(i)})^2
\]  

(20)

Thus \( \gamma_k^{(i)} \geq 0 \) and \( \sum_{k=1}^{N} \gamma_k^{(i)} = 1 \). Therefore, we may regard \( \gamma_k^{(i)} \) as the probability distribution defined by \( X_i^{(i)} \).

Suppose that the \( (k, l) \) element of symmetric matrix \( G^{(i)}(i = 1, 2) \) is

\[
g_{kl}^{(i)} = \frac{1}{N_i} \sum_{j=1}^{N_i} x_{jk}^{(i)} x_{jl}^{(i)}
\]  

(21)

Then \( \gamma_k^{(1)} = g_{k1}^{(1)} \) That is to say, the diagonal element of \( G^{(i)}(i = 1, 2) \) is \( \gamma_k^{(i)} \) defined as above. Let

\[
A = G^{(1)} - G^{(2)}
\]  

(22)

Write \( \gamma^{(i)} = (\gamma_1^{(i)}, \gamma_2^{(i)}, \ldots, \gamma_n^{(i)}) \), and every component of \( \gamma^{(i)} \) is an element of \( G^{(i)}(i = 1, 2) \) in diagonal line. Let

\[
s = s(\gamma^{(1)}, \gamma^{(2)}) = \sum_{i=1}^{n} (\gamma_i^{(1)} - \gamma_i^{(2)})^2
\]  

(23)

Then let the eigenvalues of \( A \) and corresponding eigenvectors be \( \lambda_k \) and \( u_k \), \( k = 1, 2, \ldots, n \), and we make feature extraction transform as follows

\[
Y = T'X
\]  

(24)

where \( T = (u_1, u_2, \ldots, u_d) \) is the feature extraction matrix. From here, we can reach the goal of compressing the data information.

According to the discussion above, we can get Theorem 2 as follows.

Theorem 2. Suppose that the matrix \( A \) and the function \( s \) are defined by formulæ (22) and (23), then IPE = maximum iff the coordinate system which makes

Proof Copy
2.4.2. Feature extraction effect. In order to explain the effect of feature extraction, we can use the eigenvalues \(\lambda_i (i = 1, 2, \ldots, n)\) of matrix \(A\), and arrange them as follows:

\[\lambda_1^2 \geq \lambda_2^2 \geq \ldots \geq \lambda_d^2 \geq \ldots \geq \lambda_n^2\]  \hspace{1cm} (25)

The total variance sum of the square is denoted by

\[V = \sum_{k=1}^{N_1} \lambda_k^2 = s_0\]

and then the variance square ratio is

\[V_i = \frac{\sum_{i=1}^{d} \lambda_i^2}{s_0}\]  \hspace{1cm} (26)

It can measure the degree of information compression. Generally speaking, so long as \(V_i \geq 80\%\), we can reach the goal of feature compression.

2.4.3. Algorithm. According to the discussion above, we can get the algorithm of feature extraction based on IPE as follows.

**Step 1.** Calculate the square normalization matrix for the original data of the two classes \(C_1\) and \(C_2\), and get square normalization data matrix \(X^{(1)}\) and \(X^{(2)}\).

**Step 2.** Calculate the centralization matrix for data matrix \(X^{(1)}\) and \(X^{(2)}\), and then calculate the symmetric matrix \(G^{(1)}\) and \(G^{(2)}\), then calculate difference matrix \(A\).

**Step 3.** Calculate all eigenvalues of matrix \(A\) and all eigenvectors corresponding to eigenvalues.

**Step 4.** Arrange the square of eigenvalues \(\lambda_i (i = 1, 2, \ldots, n)\) of matrix \(A\) based on formula (25), and then decide the feature extraction effect according to formula (26). Generally speaking, when \(V_i \geq 80\%\) or bigger, we can select eigenvectors \(u_1, u_2, \ldots, u_d\) corresponding to the former \(d\) eigenvalues, and construct the information extraction matrix \(T = (u_1, u_2, \ldots, u_d)\).

**Step 5.** Compress data matrix \(X^{(1)}\) and \(X^{(2)}\) based on transformation (24), and so we reach the goal of optimal information feature extraction.

### 2.5. Analysis of an example

The original data sets come from the mean temperature of the Spanish weather in 1989 (taken from the Spanish Statistics Yearbook, 1990). Among them, the first class \(C_1\) shows the monthly mean lower temperature of each inland city, in comparison with other cities of the inland district. The second class \(C_2\) means the monthly mean higher temperature of each coastal area city, in comparison with other cities in the coastal area [18].

Applying the DPS data processing system, we can get total variance \(G^{(1)}\) and \(G^{(2)}\), then get the difference matrix \(A = G^{(1)} - G^{(2)}\). In order to process the data conveniently, and avoid affecting operational accuracy, we multiply by 1000 for every element of matrix \(A\), and then we get the symmetry matrix 1000 \(A\). Finally, we calculate all the eigenvalues \(\lambda_1, \lambda_2, \ldots, \lambda_{12}\) and

### Table 1

<table>
<thead>
<tr>
<th>City</th>
<th>Class C1</th>
<th>Class C2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(y_1^{(1)})</td>
<td>(y_2^{(1)})</td>
</tr>
<tr>
<td>Lugo</td>
<td>-0.21184</td>
<td>-0.31616</td>
</tr>
<tr>
<td>Avila</td>
<td>-0.14838</td>
<td>-0.42484</td>
</tr>
<tr>
<td>Burgos</td>
<td>-0.17783</td>
<td>-0.40739</td>
</tr>
<tr>
<td>Leon</td>
<td>-0.16555</td>
<td>-0.39693</td>
</tr>
<tr>
<td>Salamanca</td>
<td>-0.17101</td>
<td>-0.39316</td>
</tr>
<tr>
<td>Soria</td>
<td>-0.15268</td>
<td>-0.41626</td>
</tr>
<tr>
<td>Guadalajara</td>
<td>-0.15811</td>
<td>-0.42626</td>
</tr>
<tr>
<td>Teruel</td>
<td>-0.16647</td>
<td>-0.39021</td>
</tr>
<tr>
<td>Vitoria</td>
<td>-0.21233</td>
<td>-0.35404</td>
</tr>
<tr>
<td>Izana</td>
<td>0.07477</td>
<td>-0.41782</td>
</tr>
</tbody>
</table>
Supervised feature extraction algorithm

corresponding eigenvectors \( u_1, u_2, \ldots, u_{12} \). According to formula (26), we select the two eigenvectors \( u_1, u_2 \), and then we construct the feature extraction matrix \( T = (u_1, u_2) \). By feature extraction transformation \( Y^{(i)} = T'X^{(i)} \) \((i = 1, 2)\), we can get the feature extraction data matrix of the matrix \( X^{(1)} \) and \( X^{(2)} \), denoted by \( Y^{(1)} \) and \( Y^{(2)} \), and listed in Table 1.

Therefore, by compressing the original 12-dimensional pattern vector, we can get a two-dimensional pattern vector loaded with 99.14% information contents of the original data. The compressed results can be seen in Figures 1 and 2. From Figures 1 and 2, we can find that the distribution of feature vectors after being compressed for class \( C_1 \) and the class \( C_2 \) is concentrated, meanwhile the within-class distance is small, the between-class distance is big, and class separability is maximum.

In order to explain the compressed effect by the algorithm here, we apply PCA compression algorithm to class \( C_1 \) and class \( C_2 \) respectively, and the results can be seen in Figures 3 and 4. We can see that the distribution of feature vectors after being compressed for class \( C_1 \) and class \( C_2 \) is one of decentralization, contrary to Figures 1 and 2. So the experimental results demonstrate that the algorithm presented here takes full advantage of the class-label information of the training samples, and is a valid and reliable feature extraction algorithm.

3. Conclusions

According to the definitions and properties of Shannon entropy and cross entropy, we have studied and
discussed the feature extraction problem in this paper. We have proved that the SCE is a kind of distance measure, and proposed a new concept of IPE based on PE, which is equivalent to SCE and can be used as the criterion of class separability. Based on IPE, we have designed a new supervised feature extraction algorithm, and applied it in practice. The experimental results show that the algorithm presented here is valid, and the compression effect is significant.

4. Acknowledgements

We wish to thank the National Natural Science Foundation of China for grant no. 60435010, the National Grand Fundamental Research (973) Program of China for grant no. 2003CB317004, the China Post-doctoral Science Foundation for grant no. 2005037439 and the Doctoral Science Foundation of Shandong Agricultural University for grant no. 23241, all supporting our research. In addition, we would like to thank the anonymous reviewers for their constructive advice.

References