International Journal of Intelligent Technologies and Applied Statistics
Vol.5, No.1 (2012) pp.101–108

Classification in Image Recognition by Ambiguous Components

Jengnan Tzeng*

Department of Mathematical Sciences, National Chengchi University, Taipei, Taiwan

ABSTRACT

In the image recognition field, there are many proposed artificial intelligence techniques for finding features that can differentiate data belonging to different classes. Features or components which appear ambiguous for separating data belonging to different classes are usually left out in this field. In this paper, we will demonstrate that by proper design those ambiguous components can still be used for differentiating data. We proposed an association rules based method for designing an image classifier that can distinguish natural images and text images. Our experiments indicate that when existing approaches fail to carry out correct classification, our method can undoubtedly achieve better results.

Keywords: Association rules; Image recognition and ambiguous region

1. Introduction

Since digital images can be regarded as high dimensional data, a common process in image recognition techniques is to look for features with distribution that would signal the distinctiveness among different types of images. For example, the supported vector machine (SVM) [1] method looks for two parallel hyper-planes that would make data belonging to two classes separated at the opposite sides of these two parallel planes. The normal vector of these two planes can be considered as a feature such that data projected to this direction have two clearly separated distributions.

In linear discriminant analysis (LDA) [2], we look for some eigenvectors such that the projections of data from two different classes are separated clearly in the subspace generated by these eigenvectors. Hence, these eigenvectors are features with less ambiguity in this problem. Either in SVM or LDA, it is possible for two classes of data to be highly mixed. Thus, it is hard to find features that can separate

* Corresponding author: jengnan@math.nccu.edu.tw

these two data well. However, when employing only these features for recognition problem, much valuable information might be lost in the process. Therefore we intend to utilize these features that are usually left out in recent artificial intelligence techniques. Upon these features we will develop an association rule based recognition method that performs acceptably.

In the direct LDA algorithm for high-dimensional data, such as image recognition data, the LDA algorithm encounters several difficulties [3]. A low resolution image of size 64-by-64 implies a feature of $64 \times 64 = 4{,}096$ dimensions, and therefore the scatter matrices of the LDA eigenvalue problem has the size of $4{,}096 \times 4{,}096 = 16$M. The first challenging is to handle such big size of matrix. If the image samples are less than 4,096, the scatter matrices are always singular. Although there are many articles to handle the singular problem in LDA [4], the main eigenvector for image recognition is derived form the combination of all the components of images. When the number of component is large, it is difficult to select the significant component from the LDA eigenvector for the classification. If we use the weighting of each component in the main eigenvector as the significance of component, we will see that the number of significant component is very sensitive to the threshold.

An image recognition problem is also related to variable and feature selection problem. The number of variables that researchers meet is increasing rapidly. Before 2,000 only few problems used more than 40 features. In recent years, most papers explore problems with hundreds to tens of thousands of variables [5]. The key challenge is the number of variables is much larger than the number of samples. For example, in the gene selection problem of microarray data, the number of variables in the range from 6,000 to 60,000, however the number of sample is in the range from 3 to 300. In the text classification problem, the vocabularies of hundreds of thousands of words are very common. If we also use large document collections for research, the SVM or LDA base method become infeasible. To generate such a big matrix of square size about hundreds of thousands and solve the eigenvalue of this huge matrix is impossible.

Our purpose is to design a new classification method that satisfies large variables and small size data. Because the data size is small, unlike some proposed method obtain the classifier by averaging or cancelling some variables, we try to use all possible data to provide the ability of recognition.

2. Definition of ambiguous components

We begin with some mathematical notations to be used. Assume data $X \in R^p$ has K classes. In this paper, we are only concerned with the case of $K = 2$. When the case of $K = 2$ is understood well, the conclusion can be readily extended for $K \geq 3$. The set of the k-th class data is denoted by C_k. Each element in C_k is denoted by $X_{k,n}$ for $n = 1,\ldots, N_k$, where N_k is the number of elements in C_k. Let F be a

transformation that maps $X_{k,n}$ from R^p to R^r. We called $F(X_{k,n})$ the feature of $X_{k,n}$. Because $F(X_{k,n}) \in R^r$, we denoted the i-th component of $F(X_{k,n})$ by $f^i_{k,n}$, and we use f^i to indicate the i-th component of R^r. Given the index k and i, the average and the standard deviation of $f^i_{k,n}$ for $n = 1,\ldots, N_k$ are denoted by μ^i_k and σ^i_k, respectively. For a given parameter λ, if

$$\frac{(\mu^i_{k_1} - \mu^i_{k_2})^2}{(\sigma^i_{k_1})^2 + (\sigma^i_{k_2})^2} > \lambda, \tag{1}$$

for some k_1 and k_2, then f^i is defined as an **efficient component** for λ in R^r. If λ is given and for all the pairs of k_1 and k_2 the above inequality is not satisfied, then f^i is defined as an **ambiguous component** for λ in R^r. It is obvious that if f^i is an ambiguous component for a small λ, there exist some classes, say k_1 and k_2, whose data are highly mixed in f^i component. We note that this component is not usually utilized in recognition techniques. Now, we define the ambiguous region as the subset of the indexes by

$$AR(\lambda) = \left\{ i \in \{1,\ldots, r\} \middle| \frac{(\mu^i_{k_s} - \mu^i_{k_t})^2}{(\sigma^i_{k_s})^2 + (\sigma^i_{k_t})^2} \le \lambda,\ \forall k_s, k_t \right\}. \tag{2}$$

It is trivial that if $\lambda_1 > \lambda_2$, we have $AR(\lambda_2) \subseteq AR(\lambda_1)$. A lot of important information in the ambiguous region will be discussed in the next section.

3. Useful interval in ambiguous component

Let $i \in AR(\lambda)$ for some small λ and assume that f^i_{l,k_s}, f^i_{l,k_t} are random variables with normal distribution. That is $f^i_{l,k_s} \sim N(\mu^i_{k_s}, \sigma^i_{k_s})$, $f^i_{l,k_t} \sim N(\mu^i_{k_t}, \sigma^i_{k_t})$ and $(\mu^i_{k_s} - \mu^i_{k_t})^2 \le \lambda\left[(\sigma^i_{k_s})^2 + (\sigma^i_{k_t})^2\right]$. Because λ is small, the overlapping region of these two distributions $N(\mu^i_{k_s}, \sigma^i_{k_s})$ and $N(\mu^i_{k_t}, \sigma^i_{k_t})$ is the high probability region of each distribution. If a testing image T has value x_0 in f^i component, then we are interested in the probability $P(T \in C_k \middle| f^i = x_0)$ for $k = 1,\ldots, K$. By Bayesian theorem, we have

$$P(T \in C_k \middle| f^i = x_0) = \frac{P(f^i = x_0 \middle| T \in C_k) P(T \in C_k)}{\sum_{j=1}^{K} P(f^i = x_0 \middle| T \in C_j) P(T \in C_j)}. \tag{3}$$

We can see that if $P(f^i = x_0 | T \in C_j) \cong 0$ for $j \neq k$ and $P(f^i = x_0 | T \in C_k) \neq 0$, $f^i = x_0$ becomes an important learning rule to identify $T \in C_k$. The learner $f^i = x_0$ could be extended to $f^i \in S$ for some subset $S \subset R$. We define a new set S_k^i by

$$S_k^i = \left\{ x_0 \in R \middle| P(X \in C_k | f^i = x_0) > P(X \in C_j | f^i = x_0), \forall k \neq j \right\}. \quad (4)$$

Then $f^i \in S_k^i$ is another learner for component f^i to identify whether an image (or object) belongs to C_k.

Theorem 1.

$S_{k_1}^i \cap S_{k_2}^i = \varnothing$ for $k_1 \neq k_2$.

The proof of this theorem is straightly forward. If $x_0 \in S_{k_1}^i$, then $P(X \in C_{k_1} | f^i = x_0) > P(X \in C_{k_2} | f^i = x_0)$ and $x_0 \notin S_{k_2}^i$.

According to the definition of S_k^i, this set might contain some isolated point, say x_1. If given any $\delta > 0$, there exists $x_2 \in S_j^i$ for some $j \neq k$, then this discontinuity does not make x_1 a good learner. To increase the stability and accuracy of prediction, we only consider the case of $S = \bigcup I_l$, where I_l is an interval contained in S_k^i and $|I_l| > 0$ for all l. Then we define $\hat{S}_k^i \subseteq S_k^i$ as the maximal set of this form, that is $\hat{S}_k^i = \bigcup_{|I_l>0,\, I_l \subseteq S_k^i|} I_l$. Similarly, we have $\hat{S}_{k_1}^i \cap \hat{S}_{k_2}^i = \varnothing$ for $k_1 \neq k_2$.

Given an ambiguous component f^i, there might exist more than one $S_k^i \neq \varnothing$ for index k. We are concerned with the quantities of $P(X \in C_{k_1} | f^i \in \hat{S}_{k_1}^i)$ and $P(X \in C_{k_2} | f^i \in \hat{S}_{k_1}^i)$, as this would decide if our classification works in this case or not. From Bayesian theorem, we know that the denominators of these two probabilities are the same, therefore we only have to compare $P(X \in C_{k_1}, f^i \in \hat{S}_{k_1}^i)$ and $P(X \in C_{k_2}, f^i \in \hat{S}_{k_1}^i)$. Because $P(X \in C_{k_1}, f^i \in \hat{S}_{k_1}^i)$ can be expressed as

$$P(X \in C_{k_1}, f^i \in \hat{S}_{k_1}^i) = \sum_{x_0 \in \hat{S}_{k_1}^i} P(X \in C_{k_1}, f^i = x_0),$$

we can easily conclude that $P(X \in C_{k_1} | f^i \in \hat{S}_{k_1}^i) \geq P(X \in C_{k_2} | f^i \in \hat{S}_{k_1}^i)$. Hence S_k^i is a useful set to identify whether $X \in C_{k_1}$ for the f^i component.

4. Combining association rules to increase the prediction accuracy

Given an image $X \in C_k$, when the f^i component is in the ambiguous region, the probability of $P(X \in C_k, f^i \in \hat{S}_k^i)$ is low. Therefore, applying only one rule to make a decision is somehow risky. If the number of non-empty sets $\{S_k^i | S_k^i \neq \varnothing, i = 1,\ldots, r\}$ is L_k, it is likely that these L_k rules have certain relationships among them. Therefore

it makes intuitive sense to combine the association rule method to increase the prediction accuracy.

The purpose of association rule [3] is to establish the relationship between a combination of input variables and a combination of output variables. Here, we give a brief introduction to the association rules method.

Let $I = i_1,\ldots, i_k$ be a set of k elements, called "items." Then, a basket data $B = b_1,\ldots, b_n$ is any collection of n subsets of I, and each subset $b_i \subseteq I$ is called a "basket" of items. We say that there is an association rule *A(Antecedent)→B(Consequent)* if:

A and B occur together in at least $s\%$ of the n baskets **Support**.

All the baskets containing A, at least $c\%$ also contains B **Confidence**. $X \in C_k$

To apply association rule to our method, we can consider each feature vector in R^r to be a basket and the learner for $X \in C_k$ as the items. We search in the training data for the association rules with significant support and confidence. By using these combinations of rules, we expect to obtain results more accurate than by the original learning rule.

5. Experimental results

We will now apply our method in the image recognition problem. There are two types of images in our application: one is the text images that are snapped from books and the other is the natural images that are snapped from the landscapes. For each type we collect 500 images; therefore in total we have 1,000 images. For convenience, we label the set of text images as C_1, and the set of natural images as C_2. We randomly choose 250 images from each of these two types of images respectively to form the training set, and the remainders will be the testing set. Because the images are not always the same size, we will rescale them to 640×480 before any further processing. Moreover, the illumination of each image is not generally the same, so we adjust the minimal illumination to zero and the maximal illumination to 255 by applying a linear transformation.

The goal is to construct a model that can identify the input image to be either a natural image or a text image. Since in text image there are usually a lot of symbols and letters, and these symbols and letters have sharp corners in the boundary of font, therefore the frequency domain is considered better for the presentation of the data than the physical domain. After we rescale the image and adjust its illumination, we apply 2D discrete cosine transform (DCT) to obtain the DCT coefficient of each image. Because the excessively high frequency is usually considered as a noise, we only extract the first 300×300 DCT coefficients as our domain, i.e., the dimension of R^r is 90,000.

In the DCT domain, we check whether the component is ambiguous. There are 86.35% of components that are ambiguous for $\lambda = 0.2$. This data matches closely to our assumption.

As there are 600 training data, we have at most 600 values in f^i for $i = 1,..., 90000$. For a fixed f^i, and for each value of f^i , say x_0, we can compute $P(X \in C_k \mid f^i = x_0)$ by

$$P(X \in C_k \mid f^i = x_0) = \frac{\#\left\{X \in C_k \mid f^i(X) = x_0\right\}}{\#\left\{X \in \bigcup_j C_j \mid f^i(X) = x_0\right\}}, \text{ for } k = 1 \text{ and } 2. \quad (5)$$

After we compute $P(X \in C_k, f^i = x_0)$, we obtain S_k^i by its definition. Due to the limited number of the training data, it is not easy to determine the interior interval of S_k^i. Hence we compute the mean and the standard deviation of $x_0 \in S_k^i$, and use the probability density function to determinate the region of interior interval of S_k^i. We then obtain $\hat{S}_k^i$, and we can further compute the learner $P(X \in C_k, f^i \in \hat{S}_k^i)$. Figure 1 is the distribution of the number of text images whose component f^i falls into $\hat{S}_k^i$. Figure 2 is the distribution of the number of natural images. If the pixel is dark blue, then the corresponding $\hat{S}_k^i$ is empty. The brightness is proportional to the number of elements in $\hat{S}_k^i$. From Figure1 and Figure 2, we can see that these two images are complementary and the significant frequency of text image is in high frequency band. This matches the results from general experiments.

Now, we can apply association rules method to observe the association rules for these two types of images. There are top six significant rules (Support > 18.2% and Confidence > 29.6%) of text image and top six significant rules (Support > 26.8% and

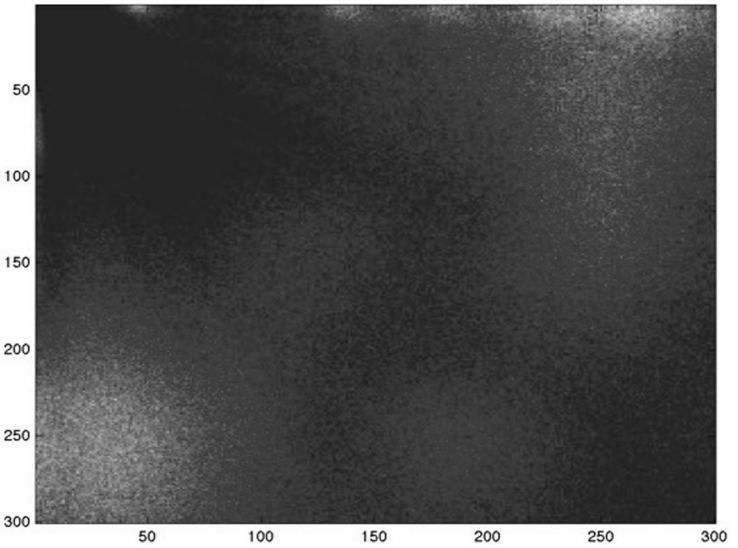

Figure 1. The distribution of ambiguous rules in the DCT base of text image (T).

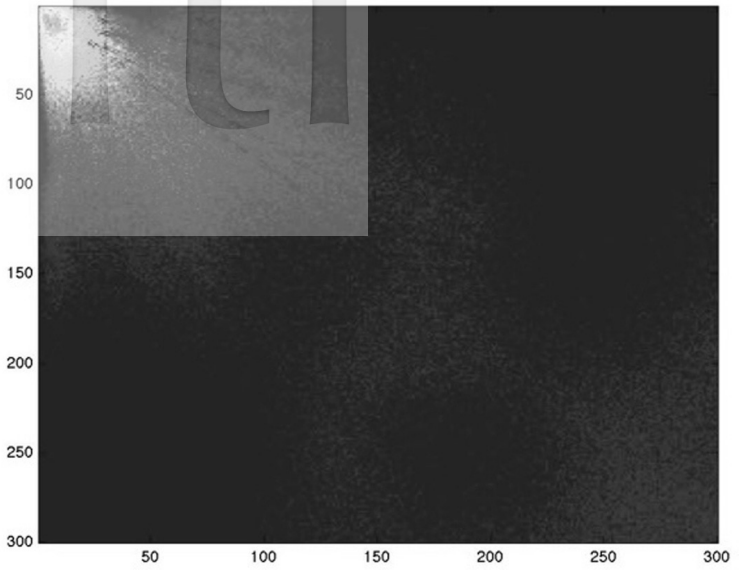

Figure 2. The distribution of ambiguous rules in the DCT base of natural image (N).

Confidence > 50%) for natural images. See Table 1. There is only one association rule that has confidence more than 95%. Most of rules are with confidence less than 60%. This shows that the ambiguity assumption fits our experimental material, namely, it is highly mixed.

Table 1. The top six rules of confidence and support for text (T) images and natural (N) images.

Confidence (T)	96.4%	42.8%	34.8%	33.6%	33.0%	29.6%
Support (T)	50.6%	22.6%	21.4%	19.0%	18.2%	18.2%
Confidence (N)	91.6%	55.6%	50.8%	50.8%	50.4%	50.0%
Support (N)	47.0%	29.8%	26.8%	28.6%	27.8%	28.2%

We use the remaining 250 images as the testing set. Each image in the testing set is also scaled in both size and illumination and then transformed to DCT frequency. For each image, we observe how many rules would pass in text image and natural image. We count the number, and assign the image type to whom the maximal number occurs. The accuracy rates of both two classes of images utilizing the top six rules are 100%. Even though we remove the top one of rules in text image and natural image, the accuracy rates are still 100%, which is quite remarkable.

6. Conclusions

In this paper we define the condition for a variable to be ambiguous. While it is generally not easy to obtain a good classification by the existing approaches if the data to be classified has many ambiguous components, our approach is able to significantly enhance the accuracy of classification. By integrating the concept of association rules method, our proposed method can attain remarkable results for image recognition using only few rules. With its characteristic, we expect to see this method to be applied successfully in many other artificial intelligence related fields.

References

[1] N. Cristianini and J. Shawe-Taylor (2000). *An Introduction to Support Vector Machines: And Other Kernel-Based Learning Methods*, Cambridge University Press, Cambridge, UK.

[2] G. J. McLachlan (2004). *Discriminant Analysis and Statistical Pattern Recognition*, Wiley, New York.

[3] H. Yu and J. Yang (2001). A direct LDA algorithm for high-dimensional data -- With application to face recognition, *Pattern Recognition*, 34, 2067-2070.

[4] M. Wang, A. Perera and R. Gutierrez-Osuna (2004). Principal discriminants analysis for small-sample-size problems: Application to chemical sensing, *Proceedings of the IEEE Sensors*, 2, 591-594, Vienna, Austria.

[5] I. Guyon and A. Elisseeff (2003). An introduction to variable and feature selection, *Journal of Machine Learning Research*, 3, 1157-1182.

[6] R. Agrawal, T. Imielinski and A. Swami (1993). Mining association rules between sets of items in large databases, *Proceedings of the ACM SIGMOD International Conference on Management of Data*, 207-216, Washington, DC.