2011 International Conference on Network and Electronics Engineering
IPCSIT vol.11 (2011) © (2011) IACSIT Press, Singapore

Classification on Ambiguous Components via Association Rules

Jengnan Tzeng and Yen-Hung Chen

Department of Mathematical Sciences, National Chengchi University

Abstract. In the image recognition field, there are many proposed artificial intelligence techniques that try to find features such that data belonging to different classes can be separated by these features. Some features or components that make ambiguous to separate data belonging to different classes are usually omitted in this field. In this paper, we will demonstrate that those ambiguous components also have distinguishable ability. We proposed an association rules based method to design an image classifier that can distinguish natural images and text images. It is no doubt that the distinguishable ability utilizing those ambiguous components is even better than traditional methods, and is also required when recent techniques meet fault.

Keywords: association rules, image recognition and ambiguous region

1. Introduction

Because digital image could be considered as a high dimensional data, a common process in image recognition technique is to look for some features that images belonging to different classes are very different in these features. For example, the Supported vector machine (SVM) [2] method looks for two parallel hyper-planes that makes data belonging to two classes ~~are~~ separated at the opposite sides of these two parallel planes. The normal vector of these two planes can be considered as a feature that data projected to this direction have two clear separated distributions.

In Linear discriminant analysis (LDA) [1], we look for some eigenvectors such that the projection of two class data are separated clearly in the subspace generated by these eigenvectors. Hence, these eigenvectors are features with less ambiguousness in this problem. Either in SVM or LDA, two classes of data highly mixed together are possible. Thus, it's hard to find features that can well separate these two data. Only using these features for recognition problem will lose much valuable information. We try to keep these features that usually are omitted in recent artificial intelligent techniques. Upon these features, we will develop an association rule based recognition method that performs acceptably.

For convenient representation, we define some mathematical notations first. Assume data $X \in R^p$ has K classes. In this paper, we only concern the case for $K = 2$. When the case of $K = 2$ is well known, the other case for $K \geq 3$ could be easily extended. 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 $n = 1,\cdots,N_k$, 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 $n = 1,\cdots,N_k$ is denoted by μ^i_k and σ^i_k, respectively. For a given parameter λ, if

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

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 are data for some k_1 and

k_2 classes highly mixed together in f^i component. Then this component is usually not utilized by recognition techniques. Now, we define the ambiguous region as the subset of the index by

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

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.

2. 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 from 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, we are interested in the probability $P(T\in C_k|f^i=x_0)$ for $k=1,\cdots,K$. By Bayesian theorem, we have

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

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 learner 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^i_k as

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

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

Theorem $S^i_{k_1}\cap S^i_{k_2}=\varnothing$ for $k_1\neq k_2$.

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

From the definition of S^i_k, this set might contain some isolated point, say x_1. If given any $\delta>0$, there exists $x_2\in S^i_j$ for some $j\neq k$, then this discontinuity makes x_1 not convincible as a good learner. For the reason of increasing 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^i_k and $|I_l|>0$ for all *l*. Then we define $\hat{S}^i_k\subseteq S^i_k$ as the maximal set of this form, that is $\hat{S}^i_k=\bigcup_{|I_l|>0,I_l\subseteq S^i_k}I_l$. Similarly, we have $\hat{S}^i_{k_1}\cap\hat{S}^i_{k_2}=\varnothing$ for $k_1\neq k_2$.

Given an ambiguous component f^i, there might exist more than one $S^i_k\neq\varnothing$ for index *k*. We concern the quantities of $P(X\in C_{k_1}|f^i\in\hat{S}^i_{k_1})$ and $P(X\in C_{k_2}|f^i\in\hat{S}^i_{k_1})$. From Bayesian theorem, we know that the denominators of these two probabilities are the same. We only have to compare $P(X\in C_{k_1},f^i\in\hat{S}^i_{k_1})$ and $P(X\in C_{k_2},f^i\in\hat{S}^i_{k_1})$. Because $P(X\in C_{k_1},f^i\in\hat{S}^i_{k_1})$ can be expressed as

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

We can easily prove that $P(X \in C_{k_1} \mid f^i \in \hat{S}^i_{k_1}) \geq P(X \in C_{k_2} \mid f^i \in \hat{S}^i_{k_1})$. Hence S^i_k is a useful set to identify whether $X \in C_{k1}$ for the f^i component.

3. Combine 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 $P(X \in C_k, f^i \in \hat{S}^i_k)$ is low. Therefore, using one rule to make a decision is somehow risky. If the number of none empty $\left\{ S^i_k \middle| S^i_k \neq \varnothing, i = 1, \cdots, r \right\}$ is L_k, these L_k rules might have some relationship between each other. It is natural to combine the association rule method to increase prediction accuracy.

The goal 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 $I = i_1, \cdots, i_k$ be a set of *k* elements called 'items'. Then, a basket data $B = b_1, \cdots, b_n$ 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**. $X \in C_k$

Applying association rule to our method, we can consider that each feature vector in R^r is a basket and the learner for $X \in C_k$ as the items. We search for the association rules from the training data such that the rules have significant support and confidence. Using these combinations of rules, we expect that we can obtain the accuracy result than the original learner.

4. Experimental results

We apply our method in the image recognition problem. There are two types of images, one is the text image that is snapped from books and the other is the natural image that is snapped from the landscapes. For each type of image, we obtain 500 images; in total we have 1000 images. For convenience, we label the text image as C_1, and the natural image as C_2. We randomly choose 250 images from these two types of images respectively to form the training set and the remainders as the testing set. Because the images are not always the same size, we will rescale them to 640 x 480 before we do further processing. Moreover, the illumination of each image is not the same generally, so we adjust the minimal illumination to zero and the maximal illumination to 255 by a linear transformation.

The goal is to construct a model that can identify the input image to be a natural image or a text image. In text image, there should be a lot of symbols and letters. These symbols and letters are sharp in the boundary of font. So, the frequency domain is considered better 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 noise, we only extract the first 300 x 300 DCT coefficient as our domain, that is the dimension of R^r is 90000.

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 is pretty matched to our assumption.

Because there are 600 training data, we have at most 600 values in f^i for $i = 1, \cdots, 90000$. For 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

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

After we compute $P(X \in C_k \mid f^i = x_0)$, we can obtain S^i_k by its definition. Because the number of the training data is not plentiful, it is not easy to determine the interior interval of S^i_k. We compute the mean and the standard deviation of $x_0 \in S^i_k$, and use the probability density function to determinate the region of

interior interval of S_k^i. We then obtained $\hat{S}_k^i$, and we can 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 $\hat{S}_k^i$. The left one is related to the text image and the other one is a natural image. 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 Figure 1, we can see that these two images are complementary and the significant frequency of text image is in high frequency band. This matched the 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 Confidence > 50%) for natural images. See Table 1.

We use the remaining 250 images as the testing set. Each image in testing set is also scaled in both size and illumination and then transformed to DCT frequency. For each image, we see how many rules passed 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 image utilized 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 also 100%. It's amazing.

5. Conclusions

We proposed a new machine learning method that is designed by the ambiguous components. This method combine the association rules to find out the efficient rules for classfication. Few rules are used to get a good recognition result. When the data in differenct level are highly overlapping, our work has better performance than recent methods. This work is not only applied in image recognition, it can be used successfully in many artificial intelligent fields.

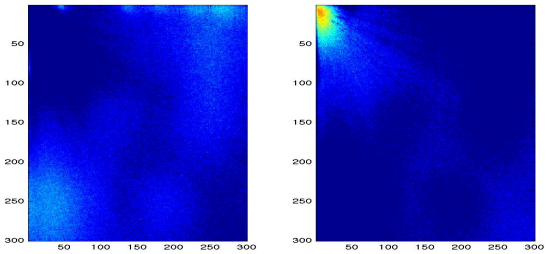

Fig. 1: The left side is the learner distribution for text image and the right side is for natural image.

Tab. 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%

6. References

[1] G. J. McLachlan. *Discriminant Analysis and Statistical Pattern Recognition*, Wiley, 2004.

[2] N. Cristianini, and J. Shawe-Taylor. *An introduction to support Vector Machines*. Cambridge University Press, 2000

[3] R. Agrawal, T. Imielinski, and A. Swami. Mining association rules between sets of items in large databases. *In Proc. Of the ACM SIGMOD Conference on Management of Data*, pages 207-216, Washington, D.C., May 1993.