ATLANTIS PRESS

Advances in Computer Science Research (ACSR), volume 76

7th International Conference on Education, Management, Information and Mechanical Engineering (EMIM 2017)

Construction of SVM Classifier for Image Retrieval

Xuejing Ding

Computer engineering college, Anhui Sanlian University, Anhui Hefei, China

330545497@qq.com

Keywords: SVM classifier; K-means clustering algorithm; Optimal selection method; Image Retrieval

Abstract. In order to make the image retrieval more quickly and efficiently, this paper proposed a new method to construct SVM classifier, it uses K-means clustering algorithm to find the representative sample in the image database, which effectively reduces the searching range of the target image, and then the optimal sample is selected from the reduced sample set as the training sample by the optimal selection method. Finally we construct the optimal training sample set which is not only large in information and low in redundancy, so as to train a better SVM classifier to get higher retrieval efficiency. The experimental results show that compared with the traditional SVM-based image retrieval method, this method can greatly improve the retrieval performance.

Introduction

At present, the mainstream image retrieval technology is content-based image retrieval (CBIR) [1]. In CBIR, there are some problems, such as the mapping between low-level image features and high-level semantic features, the subjectivity of user perception and the diversity of image content, namely the semantic gap problem. In order to solve these problems, the relevant feedback (RF) mechanism [2] is introduced, such as neural network, support vector machine SVM (Support Vector Machine)[3]. Through the study of the training sample set, we can get the corresponding model between the user's query purpose and the image feature, and then guide the new round of retrieval according to the model. The introduction of SVM greatly improves the retrieval performance of image retrieval, but how to choose the optimal training sample quickly and accurately, construct SVM classifier has become the main problem that hinders its development. The traditional SVM construction method is to select the nearest sample from the classification surface as the training sample, but this completely distance-based principle does not necessarily make the sample that selected diversity, especially if the samples in the sample set belong to a number of different categories.

This paper presents a new method based on the traditional SVM construction method, it uses K-means clustering algorithm to find the representative sample in the image database, which effectively reduces the searching range of the target image, and then the optimal sample is selected from the reduced sample set as the training sample by the optimal selection method. The training samples chosen by this method are not only close to the classification surface, but also have low redundancy among the training samples, thus ensuring the diversity of the training samples and obtaining a better classifier.

Support Vector Machine (SVM)

The basic idea of SVM is to find the kernel function and the two programming problem. Through kernel function, the data can be mapped to high dimensional feature space to solve the nonlinear problem. Radial basis function (RBF) has a wide range of convergence, which is the ideal classification basis function. Its expression is in Eq. 1:

$$K(x_i, x_j) = \exp(-\gamma\|x_i - x_j\|^2, \gamma > 0 \tag{1}$$

Where x and x_i are feature vectors, γ is the broadband of radial basis function.

In the two classification problem, assuming a given training set

$T = \{(x_1, y_1), \cdots (x_k, y_k)\} \in (R^n \times y)^k$

Where $x_i \in R^n, y_i \in Y = \{+1,-1\}, i = 1,2,\cdots,k$, and x_i is feature vector， y_i is category label. Solving of optimal classification surface will be transformed into the solving of two planning problems with constraint conditions using SVM, as shown in Eq. 2,where α_i and α_j are Lagrange Multiplier, $K(x_i, x_j)$ is kernel function.

$$\min_{\alpha} \frac{1}{2}\sum_{i=1}^{k}\sum_{j=1}^{k} y_i y_j K(x_i, x_j)\alpha_i\alpha_j - \sum_{j=1}^{k}\alpha_j \quad \text{s.t.} \quad \sum_{i=1}^{k} y_i\alpha_j = 0, 0 \le \alpha_i \le C, i = 1,2,\cdots,k \tag{2}$$

The solution of the equation is $\alpha^* = (\alpha_1^*, \cdots, \alpha_k^*)^T$ ， b^* is shown in Eq. 3:

$$b^* = y_j - \sum_{i=1}^{k} y_i\alpha_i^* K(x_i, y_j) \tag{3}$$

The optimal classification surface equation can be obtained as shown in Eq. 4:

$$f(x) = \text{sgn}(\sum_{i=1}^{k} y_i\alpha_i^* K(x_i, x) + b^*) \tag{4}$$

Paper Algorithm

K-means Clustering Algorithm. K-means clustering algorithm[4] clusters the image features of the image feature library as the initial tagged data and called representative sample.

Assuming that the number of vectors in the image feature library is n, the dimension of the vector is m, the feature vector set is D, The representative data set in D can be expressed as Tc(D),q is the query image. Using the Clustering Method to Find the Representative Data Set Tc(D) of the Training Set. In the K-means clustering method, the clustering number k is a fixed value. Set K data points (C_1, C_2,...C_k) as the center of the K cluster, In the clustering process, the clustering criterion function is used to evaluate the clustering algorithm. The clustering criterion function is defined in Eq. 5:

$$E(t) = \sum_{j=1}^{k}\sum_{x \in C_j} | x - C_j | \tag{5}$$

Where E (T) represents the sum of squared errors of all objects at the first t iteration， C_j is the jth cluster, x is the data sample used for clustering. The Euclidean distance of q to each cluster center can be expressed in Eq. 6:

$$| q - C_t | = \sqrt{(q_1 - C_{t1})^2 + \cdots + (q_{1m} - C_{1m})^2}, 1 \le t \le k \tag{6}$$

Through the K-means clustering algorithm, can find a representative data set.

Optimal Selection Method. Assume that the current optimal training set is $F = \{f_1, f_{2,} \cdots f_l\}$, The reduced sample set $U_l = \{u_{l,1}, u_{l,2} \cdots u_{l,k}\}$, The specific steps of the optimal selection method are as follows:

Calculate the distance from the sample in U_l to all the samples in F,ie

$$S_{l,i} = \sum_{j=1}^{l} d(u_{l,i}, f_j), i = 1,2,\cdots,k$$

Take the corresponding sample of U_l when $S_{l,i}$ is maximum, If there are multiple results, the sample with minimum classification distance is , recorded as $f_{l,1}$, the sample is the possible optimal training sample, and add to the possible optimal sample set: $f_{l,1} \rightarrow F_{l+1}$.

Select samples that satisfy both $| d(u_{l,i}) | < | d(f_{l,1}) |$ and $\langle u_{l,i}, f_{l,1}\rangle \ge (\pi - \langle f_{l,1}, f_j\rangle_{\min})/2, j = 1,2,\cdots,l$ from sample set U_l,add to set $U_l^{'}.U_l^{'}$ is initially empty, where $\langle f_{l,1}, f_j \rangle_{\min}$ represents the minimum sample angle for b and all samples in F.

If $U_l^{'}$ is empty, then $f_{l,1}$ is the first $l+1$ optimal training samples. Otherwise, selected the sample which have the maximum distance sum from $U_l^{'}$ to F_{l+1} and F for the possible optimal training sample, recorded as $f_{l,2}$, add to the possible optimal sample set: $f_{l,2} \to F_{l+1}$, Return to step (2).

Repeat the above steps until the sample set $U_l^{'}$ is empty, at this time $F_{l+1} = \{f_{l,1}, f_{l,1}, \cdots, f_{l,n}\}$,then $f_{l,n}$ is the first $l+1$ optimal training samples, recorded as f_{l+1} .The optimal training sample set $F = \{f_1, f_2, \cdots f_l, f_{l+1}\}$.

SVM Classifier Construction

The SVM classifier based on the above algorithm is constructed as follows:

Input: Query image Q, image database DB

Initialization: $F = \phi, DB \to U$

(1)Sort by similarity through calculating the distance between Q and the unlabeled sample U, and output a result set. If the user is satisfied with the result, the algorithm ends, otherwise go to step (2)

(2)The result set samples are labeled (related or unrelated), and the related image is added to the sample set V^+, the unrelated image is added to the sample set V^-,training set $V = V^+ \cup V^-$. SVM classifier is trained on V, the classifier h1 is obtained, and h1 is used to classify unlabeled samples to obtain positive and negative samples set U^+,U^-,$U-V = U^+ \cup U^-$.

(3)Select N images as the output of the results from $U^+ \cup V^+$ which have farthest distance from the classification surface, if the user is satisfied, the algorithm ends, otherwise go to step (4).

(4)Assume that the number of feedback samples that need to be labeled every time is K, then select the sample from U^+ as the first optimal feedback sample which is closest to the classification surface. Then use the optimal selection method to obtain the former K/2 optimal feedback samples, that is the optimal training sample set F^+.

(5)Similarly, we obtain the sample set F^- from former K/2 optimal training samples in U^-,the training result set for user labeling is $F = F^+ \cup F^-$. Then the training results are added to the training sample set: F→V, remaining sample set U-V→U, return to step (2).

Experiment and Result

In order to prove the effectiveness of the proposed algorithm, the image retrieval of the experimental in this paper and traditional SVM-based method is in the same image set to compare. The experiment selects the Corel image library, selects the representative 10 kinds of images, each class takes the 10 picture as the retrieval image.

The kernel function in the experiment uses the Gaussian kernel function in the radial basis function, ie

$$k(x, x_i) = \exp\{-(\| x - x_i \|^2)/\sigma^2\}, C = 1000, \sigma^2 = 0.5$$

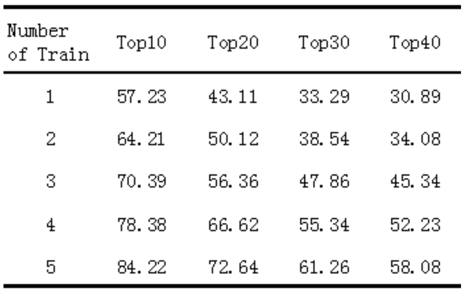
When the first search, return 10 images to the user for labeling, the training samples used for labeling will use the algorithm in this paper, the optimal sample number is 10 every time.

Using precision ratio to reflect the effect of retrieval in the experiment. Precision ratio p is the ratio of the number of images that meet the requirements(recorded as r) to the total number of images retrieved(recorded as n) in the search results, ie

p = r /n

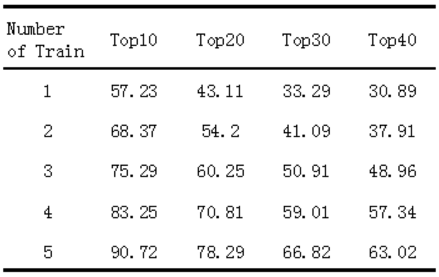
The experimental results are shown in Table 1 and Table 2:

Table 1 Retrieval performance of Traditional SVM Construction Method



Number of Train	Top10	Top20	Top30	Top40
1	57.23	43.11	33.29	30.89
2	64.21	50.12	38.54	34.08
3	70.39	56.36	47.86	45.34
4	78.38	66.62	55.34	52.23
5	84.22	72.64	61.26	58.08

Table 2 Retrieval performance of the algorithm in this paper



Number of Train	Top10	Top20	Top30	Top40
1	57.23	43.11	33.29	30.89
2	68.37	54.2	41.09	37.91
3	75.29	60.25	50.91	48.96
4	83.25	70.81	59.01	57.34
5	90.72	78.29	66.82	63.02

Experimental results show that under the same retrieval conditions compared with the traditional SVM-based method, this algorithm has an average performance improvement of more than 5%. Therefore, the method proposed in this paper can make the training sample get a better choice, get a better classifier, which greatly improve the retrieval performance.

Conclusion

This paper presents a new SVM classifier construction method. Combining the K-means algorithm with the optimal selection method so that the selected training samples not only have a large degree of information, but also have little redundancy between the training samples, so that the covered sample space can be expanded, and get better training results. Finally, the reliability and efficiency of the proposed algorithm are verified by experiments.

Acknowledgements

It is a project supported by the Natural Science Foundation of Anhui Sanlian Unversity
(No: KJZD2017010).

Reference

[1] X.Y. Li, Y.T. Zhuang and Y.H. Pan. The technique and systems of content-based image retrieval [J].Journal of Computer Research and Development,Vol.38(2001) No.3,P.344. (In Chinese)

[2] H. Wu,H.Q. Lu and S.D. Ma. A survey of relevance feedback techniques in content-based image retrieval [J].Chinese Journal of Computers,Vol.28(2005) No.12,P.1969. (In Chinese)

[3] X.J. Wang, G.C. Luo and K. Qin. Image retrieval method based on SVM and active learning [J].Application Research of Computer,Vol.33(2016) No.12,P.3836. (In Chinese)

[4] J.X. He: Research on software reliability model based on support vector machine (MS., Lanzhou University of Technology, China, 2009), P.25.(In Chinese)