Research Journal of Applied Sciences, Engineering and Technology 5(20): 4850-4856, 2013
ISSN: 2040-7459; e-ISSN: 2040-7467

Submitted: September 22, 2012 Accepted: November 12, 2012 Published: May 15, 2013

Kernel Selection of SVM for Commerce Image Classification

[1, 2]Lou Xiongwei and [2]Huang Decai
[1]College of Information Engineering, ZUT, Hangzhou, Zhejiang, China
[2]College of Information Engineering, ZAFU, Linan, Zhejiang, China

Abstract: Content-based image classification refers to associating a given image to a predefined class merely according to the visual information contained in the image. In this study, we employ SVM (Support Vector Machine) and presented a few kernels specifically designed to deal with the problem of content-based image classification. Several common kernel functions are compared for commerce image classification with the PHOW (Pyramid Histogram of visual Words) descriptors. The experiment results illustrate that chi-square kernel and histogram intersection kernel are more effective with the histogram based image descriptor for commerce image classification.

Keywords: Commerce image classification, kernel selection, SVM

INTRODUCTION

Support vector machine is a popular discriminative learning method with the advantages of providing a good out-of-sample generalization and less need of prior knowledge about the problem. In this study we employed SVM and presented a few kernels specifically designed to deal with the problem of content-based commerce image classification. The so-called content-based image classification refers to associating a given image to a predefined class merely according to the visual information contained in the image (Kannan *et al*., 2011; Wang *et al*., 2011; Perronnin *et al*., 2010; Boiman *et al*., 2008). For example, object detection, which is aimed to find one or more instances of an object in an image, is one kind of image classification problem. A second problem is view-based object recognition, the objects to be detected are instances of a certain class (e.g., apples or cars), while objects to be recognized are instances of the same object viewed under different conditions (e.g., the specially designated apple or car). A third problem is visual categorization, which refers to associating an image to two or more image categories, the former is binary classification and the latter is the so-called multiclass classification. Images depend on various parameters controlling format, size, resolution and compression quality, which make it difficult to compare image visual content, or simply to recognize the same visual content in an image saved with different parameters. This is in sharp contrast to other application domains like, for example, text categorization and bioinformatics.

The aim of this study is to select the kernels well-suited to solve commerce image classification problems with support vector machine. Kernel-based methods maps the data from the original input feature space to a high-dimensionality kernel feature space and then solve a linear problem(find a largest margin hyper-plane) in the kernel space. The Kernel-based methods allow us to interpret and design learning algorithms geometrically in the kernel space, which is nonlinearly related to the feature space. The kernel functions are firstly required to meet the so called Mercer's conditions and a well-suited kernel should incorporate the prior knowledge of the solving problem.

SUPPORT VECTOR MACHINE

Support vector machines have largely been motivated and analyzed with a theoretical framework, which is known as statistical learning theory (also called computational learning theory). Support vector machines produce nonlinear boundaries by constructing a linear boundary in a large, transformed version of the feature space and give good generalization and bounds for the computational cost of learning. In fact the SVM classifier is solving a function-fitting problem using a particular criterion and form of regularization and have some edge on the curse of dimensionality.

Hard margin SVM: Let $\{x_i, y_i, i=1,2,\ldots N\}$, x_i denote the feature vectors of the training set X and y_i is their category label, which associate either of two classes, ω_i, ω_j and are assumed to be linearly separable, as illustrated in Fig. 1. The designed hyper-plane:

$$g(x) = \omega^T x + \omega_0 = 0 \quad (1)$$

Corresponding Author: Lou Xiongwei, College of Information Engineering, ZUT, Hangzhou, Zhejiang, China

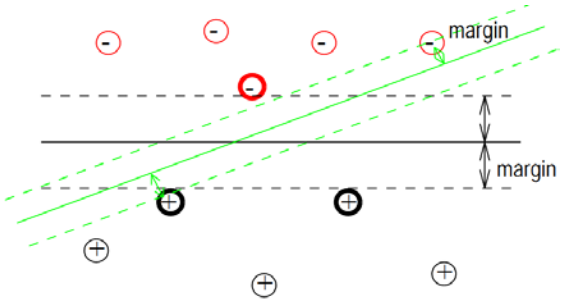

Fig. 1: The diagrammatic sketch of hard margin SVM

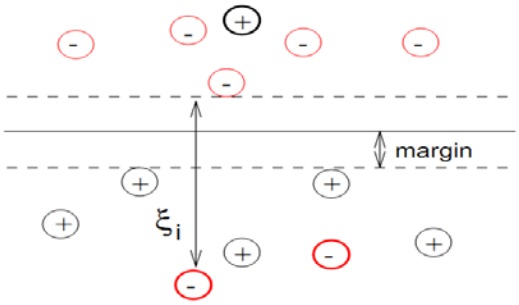

Fig. 2: The diagrammatic sketch of soft-margin support vector machine

can classify correctly all the training samples. For the largest margin hyperplane, the optimization problem is as below: Minimize:

$$\frac{\|\omega\|^2}{2} \tag{2}$$

subject to:

$$y_i(\omega.x_i + b) \geq 1 \tag{3}$$

This is the same as the original objective:

$$\frac{1}{m}\sum_{i=1}^{m} l(\omega.x_i + b, y_i) + \|\omega\|^2 \tag{4}$$

$l(y, y') = \max(0, 1 - yy')$ is the so-called hinge loss.

Soft-margin: To deal with the non-separable case, the problem is as below;
Minimize:

$$\frac{\|\omega\|^2}{2} + C\sum_{i=1}^{m} \xi_i \tag{5}$$

subject to:

$$y_i(\omega.x_i + b) \geq 1 - \xi_i \tag{6}$$
$$\xi_i \geq 0$$

This is the same as the original objective:

$$\frac{1}{m}\sum_{i=1}^{m} l(\omega.x_i + b, y_i) + \|\omega\|^2 \tag{7}$$

l(y, y`) = max (0.1- yy`)) is the so-called hinge loss.

C (capacity) is a tuning parameter to weight in-sample classification errors and it controls the generalization ability of SVM. The higher is the parameter C, the higher is the weight of in-sample misclassifications and the lower the generalization of the machine is. C is also linked to the width of the margin. The smaller is C, the larger is the margin and the more in-sample classification errors are permitted. Through Lagrange dual optimization, we construct and solve the convex programming problem:

$$\begin{aligned} \min_{\alpha} \quad & -\sum_{i=1}^{m}\alpha_i + \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j k(x_i, x_j), \\ s.t. \quad & \sum_{i=1}^{m}\alpha_i y_i = 0, \\ & 0 \leq \alpha_i \leq C, \ i = 1, 2...m; \end{aligned} \tag{8}$$

where, $k(x_i, x_j) = \Phi(x_i)^T \Phi(x_j)$ is called kernel function. The coefficients $\alpha^* = (\alpha^*_1, \alpha^*_m)^T$ are always sparse, that is, there is little number of non-zeros coefficient, which are named as support vector.

Select a coefficient $\alpha^*_j \in (0, C)$, calculate the coefficient b^*:

$$b^* = y_j - \sum_{i=1}^{m} y_i \alpha_i^* k(x_i, x_j) \tag{9}$$

and the final classification decision function is :

$$f(x) = \text{sgn}(\sum_{i=1}^{m} \alpha_i^* y_i k(x_i, x) + b^*) \tag{10}$$

Figure 2 illustrates the diagrammatic sketch of soft-margin support vector machine

Multiclass classification: The support vector machine is fundamentally a two-class classifier, however, in practice, many problems involving K>2 classes need to be tackled. In fact, the support vector machine can be extended to multiclass problems through solving many two-class problems. Several methods have been proposed for combining multiple two-class SVMs to build a multiclass classifier.

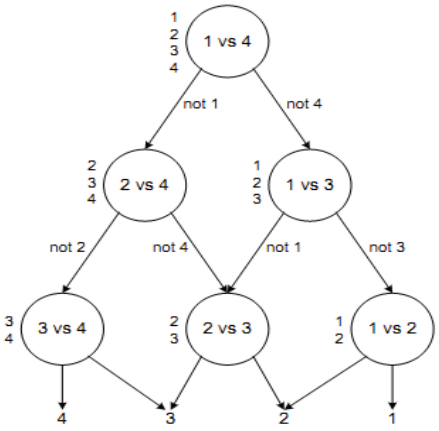

Fig. 3: The diagrammatic sketch of DAG for finding the best class out of classes

- **One-versus-the-rest approach:** One-versus-the-rest approach is one commonly used approach, it constructs K separate SVMs, in which the kth model is trained using the data from class Ck as the positive examples and the data from the remaining K-1 classes as the negative examples. However, the training sets of the one-versus-the- are imbalanced.
- **One-versus-one approach:** The one-versus-the-rest approach is to train K(K−1)/2 different 2-class SVMs on all possible pairs of classes and then to classify test points according to which class has the highest number of 'votes'. However, this approach can also lead to ambiguities in the resulting classification.
- **DAGSVM approach:** This method organizes the pairwise classifiers into a directed acyclic graph. For K classes, the DAGSVM has a total of K (K − 1)/2 classifiers and to classify a new test sample, only K − 1 pairwise classifiers need to be evaluated. The particular classifiers used depend on which path through the graph is traversed. (see the example of four-classification in Fig. 3 (Platt *et al*., 2000).

KERNEL FUNCTION FOR IMAGE CLASSIFICATION

If kernel k (,) satisfies (1) and (2), then, k (,) is a valid kernel. (Mercer kernel)

- Symmetric:

$$k(x_i, x_j) = k(x_j, x_i) \tag{11}$$

- Positive definite, that is: for all

$$\alpha \in R^N,\ \alpha^T K \alpha \geq 0 \tag{12}$$

- K is N*N gram matrix, $K_{i,j}$ = k (x_i, x_j) and it often refers to the kernel matrix.

Table 1: The kernel matrix K

	1	2	…	m
1	k_1 (x_1, x_1)	k (x_1, x_2)	…	k (x_1, x_m)
2	k (x_2, x_1)	k (x_2, x_2)	…	k (x_2, x_m)
…	…	…	…	…
m	k (x_m, x_1)	k (x_m, x_2)	…	k (x_m, x_m)

The so-called kernel matrix (gram matrix) contains all the available information for performing the learning step (Shawe-Taylor and Cristianini, 2004; Barla *et al*., 2002), which is illustrated in Table 1. Through the kernel matrix, the learning machine obtains the information about the feature space selection and the training data itself, as illustrated in Fig. 4 (Shawe-Taylor and Cristianini, 2004).

Example kernels for image classification:

- Linear kernel:

$$k(x, y) = x^T y \tag{13}$$

- Polynomial kernel: for any d>0:

$$k(x, y) = (1 + x^T y)^d \tag{14}$$

Polynomial kernel contains all the polynomial terms to the degree *d*.

- Gaussian kernels:

$$\text{for } \sigma \quad k(x, y) = \exp\left(-\|x - y\|^2 / 2\sigma^2\right) \tag{15}$$

For the Gaussian kernel, the feature space is of infinite dimension. σ is the parameter which is called 'kernel width'.

- Chi-square kernel:

$$k(x, y) = \exp(-\rho\chi^2(x, y))$$

$$\chi^2(x, y) = \sum_i \frac{(x_i - y_i)^2}{x_i + y_i} \tag{16}$$

the parameter p is often valued as the inverse of the average Chi-square distance.

- Histogram intersection kernel (Barla *et al*., 2003):

$$k_{HI}(x, y) = \sum_i \min(x_i, y) \tag{17}$$

It is a challenge task to build kernel functions, as they are not only to satisfy certain mathematical requirements but also to incorporate the prior knowledge of the application domain. For the image context, it is an extremely difficult problem, as the

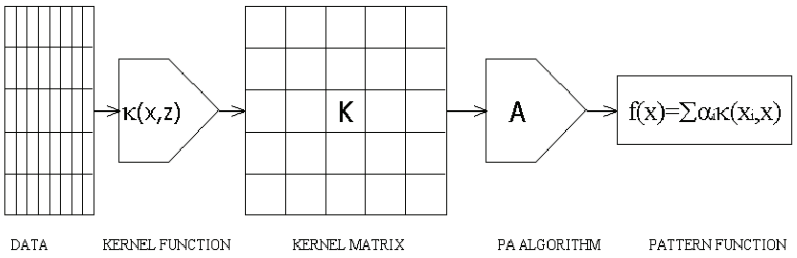

Fig. 4: Kernel matrix for pattern function

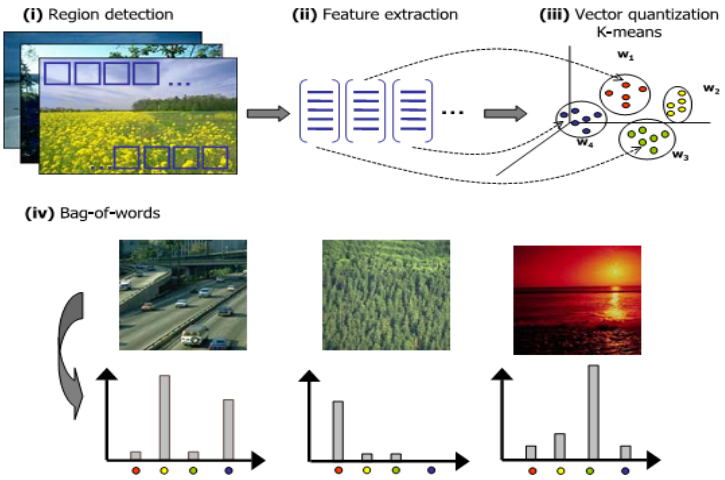

Fig. 5: The diagrammatic sketch of bag of words

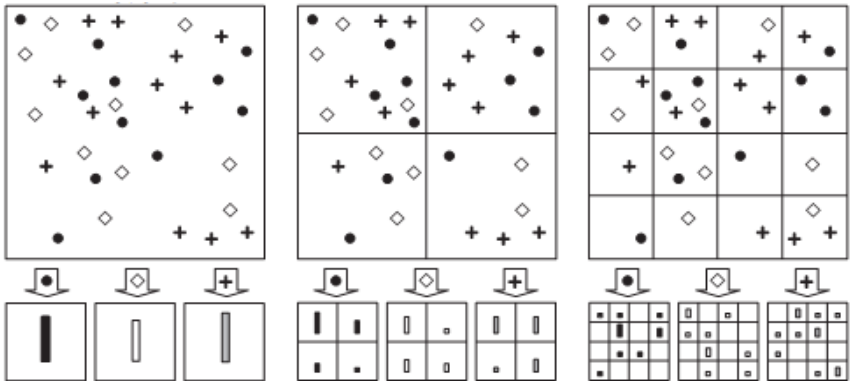

Fig. 6: Pyramid of histogram of words (PHOW)

classification has to face the large variability of the images (Chapelle *et al*., 2002).

IMAGE DESCRIPTOR

In our experiments, we employed the so-called PHOW (pyramid histogram of words) (Lazebnik *et al*., 2006) as the image descriptor. PHOW descriptor is based on the popular BOW (bag of word) model (Bosch and Marti, 2007). The basic idea of BOW model, which is borrowed from text classification, is to represent an image as a histogram of visual keywords (visual words). The visual words are actually the clustering centers of local features in the images and all the collection of the visual words are called bag of words. Figure 5 illustrates the construction of BOW, which is as follows:

- Automatically detect of interest points/interest areas or local blocks
- Represent the local areas as local descriptors (such as SIFT (Lowe, 2004; Mikolajczyk and Schmid, 2005)

Table 2: Commerce image for classification (20categories)

Bike						
Telescope						
Bird cage						
Bird house						
Opener						
Bracelet						
Handbag						
Cabinet						
Calculator						
Calendar						
Video cassette						
Tin						
Car GPS						
Handphone						
Cheese						
Chocolate						
Cigar						
Coin						
Keyboard						

Table 3: The classification accuracies of 20 commerce categories (%)

Number of training samples each category	Linear	Gaussian	Poly nominal	Histogram intersection	χ^2
5	60.2	72.1	60.3	78.6	78.5
15	65.7	78.3	66.2	82.9	83.3
30	66.2	79.8	67.5	84.8	85.6
50	67.2	82.3	67.9	85.1	85.8

- Cluster all the image descriptors with clustering algorithm (i.e., K-means) to form a number of cluster centers (visual words)
- Calculate the visual keywords distribution in an image and form the visual keywords histogram

The traditional BOW model ignores the characteristics of the spatial position of the images and employs sparse sampling mode. Lazebnik *et al*. (2006) proposed a improved descriptor named as PHOW (Pyramid Histogram Of Words). The improvements are from two aspects (Fig. 6):

- Use dense sampling instead of sparse sampling for feature extraction. The sampling interval is set to eight pixels, each 16 × 16 pixel block forms a 128-dimensional SIFT descriptor.
- Represent an image with multiple resolutions (from low resolution to high resolution), each with a series of visual keywords in the feature space. In this study, the pyramid level is set to 3 (l = 0, 1, 2) and the number of visual words is 300, then the eventually formed PHOW dimension: 300 +300 × 4 +300 × 16 = 6300.

EXPERIMENT

Experiment set: All the experiments were performed on a computer with Intel Pentium CPU 2.66GHz and 4GB RAM, which run Windows XP and MATLAB2010. The popular SVM toolbox- Libsvm (Chang and Lin, 2001) were employed. For multiple-class classification, the large margin DAG strategy (Platt *et al*., 2000) is adopted. The kernel parameters (C, σ) were obtained through a ten-fold cross-validation on each training set.

Table 2 illustrates samples of 20 commerce categories (Microsoft Research, 2010) (bike, telescope, bird cage, bird house, opener, bracelet, handbag, bracelet, cabinet, calendar, calculator, calendar, video cassette, tin, car GPS, hand phone, cheese, chocolate, cigar, coin keyboard).

RESULTS AND DISCUSSION

Table 3 illustrates the classification accuracies of 20 commerce categories with the five kernel functions. From the experiment results, we can conclude as blow:

- Chi-square kernel and histogram intersection kernel performs much better than the three general kernels (linear kernel, Gaussian kernel and poly nominal kernel).
- Chi-square kernel is slightly superior to the Histogram intersection kernel, while the linear kernel perform worst.
- The performances of all the kernels are becoming better as the number of the training samples with each category increase, particularly as the training sets increase from 5 to 15 samples each category. The average accuracies are becoming relatively stable as the training samples of each category are up to 30.

CONCLUSION

In this study, we compare several common kernel functions for commerce image classification with the PHOW descriptors. The experiments illustrate that chi-square kernel and histogram intersection kernel are more effective with the histogram based image descriptor for commerce image classification. However, it is still a challenge task to construct more appropriate kernel functions for image classification.

REFERENCES

Bosch, A.X.M. and R. Marti, 2007. Which is the best way to organize/classify images by content? Image Vision Comput., 25(6): 778-791.

Barla, A., F. Emanuele, O. Francesca and V. Alessandro, 2002. Image kernels. Lect. Notes Comput. Sc., pp: 617-628, Doi: 10.1007/3-540-45665-1_7.

Barla, A., F. Odone and A. Verri, 2003. Histogram intersection kernel for image classification. Proceeding of International Conference on Image Processing (ICIP). Italy, 3: 513-516.

Boiman, O., E. Shechtman and M. Irani, 2008. In defense of nearest-neighbor based image classification. Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp: 1-8.

Chang, C. and C. Lin, 2001. LIBSVM: A Library for Support Vector Machines. Retrieved from: http://wenku.baidu.com/view/b50dec6cb84ae45c3b358c18.html.

Chapelle, O., P. Haffner and V. Vapnik, 2002. Support vector machines for histogram-based image classification. IEEE T. Neural Networ., 10(5): 1055-1064.

Kannan, A., P.P. Talukdar, N. Rasiwasia and K. Qifa, 2011. Improving product classification using images. Proceeding of IEEE 11th International Conference on Data Mining, pp: 310-319.

Lazebnik, S., C. Schmid and J. Ponce, 2006. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. Proceeding of IEEE Computer Society Conference Computer Vision and Pattern Recognition, pp: 2169-2178.

Lowe, D., 2004. Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vision, 60(2): 91-110.

Microsoft Research, 2010. Product Image Categorization Data Set (PI 100). Retrieved from: http://research.microsoft.com/en-us /people/xingx /pi100.aspx.

Mikolajczyk, K. and C. Schmid, 2005. A performance evaluation of local descriptors. Pattern Anal. Mach. Intell., 27(10): 1615-1630.

Perronnin, F., J. Sánchez and T. Mensink, 2010. Improving the fisher kernel for large-scale image classification. Comput. Vision ECCV, 6314: 143-156.

Platt, J.C., N. Cristianini and J. Shawe-Taylor 2000. Large margin DAGs for multiclass classification. Adv. Neur. In., 12(3): 547-553.

Shawe-Taylor, J. and N. Cristianini, 2004. Kernel Methods for Pattern Analysis. Cambridge University Press, Cambridge.

Wang, Z., Y. Hu and L.T. Chia, 2011. Improved learning of I2C distance and accelerating the neighborhood search for image classification. Pattern Recogn., 44(10-11): 2384-2394.