Feature Selection for Fast Image Classification with Support Vector Machines

Zhi-Gang Fan, Kai-An Wang, and Bao-Liang Lu

Departmart of Computer Science and Engineering, Shanghai Jiao Tong University,
1954 Hua Shan Road, Shanghai 200030, China
{zgfan,kaianwang}@sjtu.edu.cn, blu@cs.sjtu.edu.cn

Abstract. According to statistical learning theory, we propose a feature selection method using support vector machines (SVMs). By exploiting the power of SVMs, we integrate the two tasks, feature selection and classifier training, into a single consistent framework and make the feature selection process more effective. Our experiments show that our SVM feature selection method can speed up the classification process and improve the generalization performance of the classifier.

1 Introduction

Pattern classification is a very active research field in recent years. As a result of statistical learning theory, support vector machines (SVMs) is an effective classifier for the problems of high dimension and small sample sets. This is a very meaningful breakthrough for machine learning and pattern classification because both high dimension and small sample set problems are too difficult to be solved by classical paradigms. According to the principle of structural risk minimization, SVMs can guarantee a high level of generalization ability. SVMs can obtain an optimal separating hyperplane as a trade-off between the quality of empirical risk and the complexity of the classifier. Furthermore, SVMs can solve linearly non-separable problems using kernel functions, which map the input space into a high-dimensional feature space where a maximal margin hyperplane is constructed [1].

In fact, SVMs are not only a good classification technique but also a good feature selection method. The problem of feature selection is well known in machine learning. Data overfitting arises when the number of features is large and the number of training samples is comparatively small. This case is very common especially in image classification. Therefore, we must find a way to select the most informative subset of features that yield best classification performance for overcoming the risk of overfitting and speeding up the classification process. By investigating the characteristics of SVMs, it can be found that the optimal hyperplane and support vectors of SVMs can be used as indicators of the important subset of features. Therefore, through these indicators, the most informative features can be selected effectively.

N.R. Pal et al. (Eds.): ICONIP 2004, LNCS 3316, pp. 1026–1031, 2004.

2 SVMs and Feature Ranking

2.1 Support Vector Machines

Support vector machine is a machine learning technique that is well-founded in statistical learning theory. Statistical learning theory is not only a tool for the theoretical analysis but also a tool for creating practical algorithms for pattern recognition. This abstract theoretical analysis allows us to discover a general model of generalization. On the basis of the VC dimension concept, constructive distribution-independent bounds on the rate of convergence of learning processes can be obtained and the structural risk minimization principle has been found. The new understanding of the mechanisms behind generalization not only changes the theoretical foundation of generalization, but also changes the algorithmic approaches to pattern recognition.

As an application of the theoretical breakthrough, SVMs have high generalization ability and are capable of learning in high-dimensional spaces with a small number of training examples. It accomplishes this by minimizing a bound on the empirical error and the complexity of the classifier, at the same time. With probability at least $1 - \eta$, the inequality

$$R(\alpha) \leq R_{emp}(\alpha) + \Phi\left(\frac{h}{l}, \frac{-\log(\eta)}{l}\right) \tag{1}$$

holds true for the set of totally bounded functions. Here, $R(\alpha)$ is the expected risk, $R_{emp}(\alpha)$ is the empirical risk, l is the number of training examples, h is the VC dimension of the classifier that is being used, and $\Phi(\cdot)$ is the VC confidence of the classifier.

According to equation (1), we can find that the uniform deviation between the expected risk and empirical risk decreases with larger amounts of training data l and increases with the VC dimension h. This leads us directly to the principle of structural risk minimization, whereby we can attempt to minimize at the same time both the actual error over the training set and the complexity of the classifier. This will bound the generalization error as in (1). This controlling of both the training set error and the classifier's complexity has allowed SVMs to be successfully applied to very high dimensional learning tasks.

We are interesting in linear SVMs because of the nature of the data sets under investigation. Linear SVMs uses the optimal hyperplane

$$(w \cdot x) + b = 0 \tag{2}$$

which can separate the training vectors without error and has maximum distance to the closest vectors. To find the optimal hyperplane one has to solve the following quadratic programming problem: minimize the functional

$$\Phi(w) = \frac{1}{2}(w \cdot w) \tag{3}$$

under the inequality constraints

$$y_i[(x_i \cdot w) + b] \geq 1, \quad i = 1, 2, \ldots, l. \tag{4}$$

where $y_i \in \{-1, 1\}$ is class label. We can obtain the functional

$$W(\alpha) = \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i,j}^{l} \alpha_i \alpha_j y_i y_j x_i^T x_j \tag{5}$$

It remains to maximize this functional under the constraint

$$\sum_{i=1}^{l} \alpha_i y_i = 0, \quad \alpha_i \geq 0, \quad i = 1, \ldots, l \tag{6}$$

Once the optimization problem has been solved, we can obtain w as follows:

$$w = \sum_{i=1}^{l} \alpha_i y_i x_i \tag{7}$$

It is usually the case that most of the parameters α_i are zero. The decision hyperplane therefore only depends on a smaller number of data points with non-zero α_i; these data points are called support vectors. So we can change the equation (7) as

$$w = \sum_{i \in SV} \alpha_i y_i x_i \tag{8}$$

As a result, equation (2) can be obtained and the SVM classifier has been built.

2.2 Feature Selection and Classification

According to the hyperplane as shown in equation (2), the linear discriminant function can be constructed for SVMs classifier as follows:

$$D(x) = (w \cdot x) + b \tag{9}$$

The inner product of weight vector $w = (w_1, w_2, \ldots, w_n)$ and input vector $x = (x_1, x_2, \ldots, x_n)$ determines the value of $D(x)$. Fig.1 shows that the $|w_k|$ of a SVMs example with R^{4096} input space has obvious variance. Intuitively, the input features in a subset of $(x_1, x_2, \ldots, x_n)$ that are weighted by the largest absolute value subset of $(w_1, w_2, \ldots, w_n)$ influence most the classification decision. If the classifier performs well, the input features subset with the largest weights should correspond to the most informative features [4]. Therefore, the weights $|w_k|$ of the linear discriminant function can be used as feature ranking coefficients. However, this way for feature ranking is a greedy method and we should look for more evidences for feature selection. In [7], support vectors have been used as evidence.

Assume the distance between the optimal hyperplane and the support vectors is Δ, the optimal hyperplane can be viewed as a kind of Δ-margin separating hyperplane which is located in the center of margin $(-\Delta, \Delta)$. According to [3],

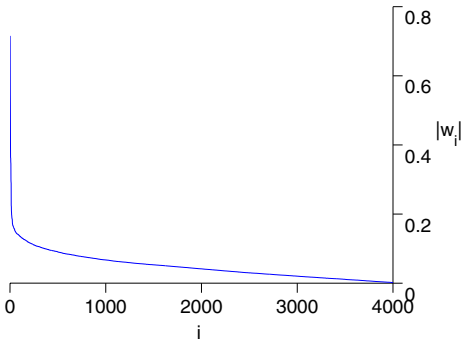

Fig. 1. $|w_i|$ ordered decreasingly in a linear SVMs example with R^{4096} input space.

the set of Δ-margin separating hyperplanes has the VC dimension h bounded by the inequality

$$h \leq \min\left(\left[\frac{R^2}{\Delta^2}\right], n\right) + 1 \tag{10}$$

where R is the radius of a sphere which can bound the training vectors $x \in X$.

Inequality (10) points out the relationship between margin Δ and VC dimension: a larger Δ means a smaller VC dimension. Therefore, in order to obtain high generalization ability, we should still maintain margin large after feature selection. However, because the dimensionality of original input space has been reduced after feature selection, the margin is always to shrink and what we can do is trying our best to make the shrink small to some extent. Therefore, in feature selection process, we should preferentially select the features which make more contribution to maintaining the margin large. This is another evidence for feature ranking. To realize this idea, we introduce a coefficient given by

$$c_k = \left| \frac{1}{l_+} \sum_{i \in SV_+} x_{i,k} - \frac{1}{l_-} \sum_{j \in SV_-} x_{j,k} \right| \tag{11}$$

where SV_+ denotes the support vectors belong to positive samples, SV_- denotes the support vectors belong to negative samples, l_+ denotes the number of SV_+, l_- denotes the number of SV_-, and $x_{i,k}$ denotes the kth feature of support vector i in input space R^n.

The larger c_k indicates that the kth feature of input space can make more contribution to maintaining the margin large. Therefore, c_k can assist $|w_k|$ for feature ranking. The solution is that, combining the two evidences, we can order the features by ranking $c_k|w_k|$. We present below an outline of the feature selection and classifier training algorithm.

- Input:
  Training examples
  $$X_0 = [x_1, x_2, \ldots x_l]^T$$

- Initialize:
  Indices for selected features: $s = [1, 2, \ldots n]$
  Train the SVM classifier using samples X_0
- For $t = 1, \ldots, T$:
  1. Compute the ranking criteria $c_k|w_k|$ according to the trained SVMs
  2. Order the features by decreasing $c_k|w_k|$, select the top M_t features, and eliminate the other features
  3. Update s by eliminating the indices which not belong to the selected features
  4. Restrict training examples to selected feature indices

$$X = X_0(:, s)$$

  5. Train the SVM classifier using samples X
- Outputs:
  The final SVM classifier and features selected by SVMs

Usually, the iterative loop in the algorithm should be terminated before the training samples can not be separated by a hyperplane. Clearly, this algorithm can integrate the two tasks, feature selection and classifier training, into a single consistent framework and make the feature selection process more effective.

3 Experiments

In order to verify the effect of our SVM feature selection method, we use the SVMs without feature selection and the SVMs with feature selection respectively in our experiments for comparison study. Two other feature selection methods (proposed in [4] , [7]) have been compared with our method. The data set used in the first experiment has totally 3433 samples which are all industrial images from a manufacturing company and 2739 samples were selected as training set, the other 694 samples were selected as test set.

In the second experiment, we use the ORL face database of Cambridge University. The non-face images (negative samples) are obtained from the Ground Truth database of Washington University and the total sample size is 2551. Table 1 and Table 2 show the test results after training. Through these results, we see that the success rate can be improved and classification speed increases rapidly at the same time in the test phase using our method.

4 Conclusion and Future Work

On the basis of statistical learning theory, we have presented a feature selection method using SVMs. Our experiments show that this method can remarkably speed up the classification process and improve the generalization performance of the classifier at the same time. In the future work, we will enhance this method and apply it to face classification.

Table 1. Test result on industrial images.

Methods	No. features	Success rate (%)	Test time (s)	Speedup
No selection	4096	96.83	69.2	-
SVM RFE in [4]	500	97.98	2.5	27.68
Selection method in [7]	500	97.55	1.8	38.44
Our method	500	98.27	2.1	32.95

Table 2. Test result on ORL face database.

Methods	No. features	Success rate (%)	Test time (s)	Speedup
No selection	10304	97.43	320.9	-
SVM RFE in [4]	4000	97.62	52.6	6.10
Selection method in [7]	4000	97.33	52.8	6.08
Our method	4000	97.71	51.5	6.23

Acknowledgements

This work was partially supported by the National Natural Science Foundation of China via the grant NSFC 60375022. The authors thank Mr. Bin Huang for the help on preprocessing the training and test data sets.

References

1. Vapnik, V. N.: Statistical Learning Theory. Wiley, New York (1998)
2. Vapnik, V. N.: The Nature of Statistical Learning Theory. Springer-Verlag, New York (2000)
3. Vapnik, V. N.: An Overview of Statistical Learning Theory, IEEE Trans. Neural Networks. vol. 10, no.5, (1999) 988-999
4. Guyon, I., Weston, J., Barnhill, S., Vapnik, V. N.: Gene Selection for Cancer Classification using Support Vector Machines. Machine Learning, vol. 46, (2002) 389-422
5. Mao, K. Z.: Feature Subset Selection for Support Vector Machines Through Discriminative Function Pruning Analysis. IEEE Trans. Systems, Man, and Cybernetics, vol. 34, no. 1, (2004) 60-67
6. Evgeniou, T., Pontil, M., Papageorgiou, C., Poggio, T.: Image Representations and Feature Selection for Multimedia Database Search. IEEE Trans. Knowledge and Data Engineering, vol. 15, no. 4, (2003) 911-920
7. Heisele, B., Serre, T., Prentice, S., Poggio, T.: Hierarchical classification and feature reduction for fast face detection with support vector machine. Pattern Recognition, vol. 36, (2003) 2007-2017