Proceedings of the Sixth International Conference on Machine Learning and Cybernetics, Hong Kong, 19-22 August 2007

FEATURE SELECTION BY COMBINING FISHER CRITERION AND PRINCIPAL FEATURE ANALYSIS

SA WANG[1], CHENG-LIN LIU[2], LIAN ZHENG[1]

[1]School of Aerospace Science and Engineering, Beijing Institute of Technology, Beijing 100081, China
[2]National Laboratory of Pattern Recognition, Institute of Automation, China Academy of Science, Beijing 100080, China
E-MAIL: swang@nlpr.ia.ac.cn, liucl@nlpr.ia.ac.cn, zheng_lian@bit.edu.cn

Abstract:

Feature selection is one of the most important issues in the fields such as data mining, pattern recognition and machine learning. In this study, a new feature selection approach that combines the Fisher criterion and principal feature analysis (PFA) is proposed in order to identify the important (relevant and irredundant) feature subset. The Fisher criterion is used to remove features that are noisy or irrelevant, and then PFA is used to choose a subset of principal features. The proposed approach was evaluated in pattern classification on five publicly available datasets. The experimental results show that the proposed approach can largely reduce the feature dimensionality with little loss of classification accuracy.

Keywords:

Feature selection; Fisher criterion; Principal feature analysis (PFA); Pattern classification

1. Introduction

Feature selection is a process of choosing a subset of preliminary features by removing irrelevant and redundant features. It is an important step in designing classification systems, and has attracted intensive attention in several fields, such as data mining, pattern recognition and machine learning, where irrelevant and redundant features are commonly encountered.

In last decades, many methods have been proposed for feature selection. John et al. [1] group them into two categories: filters and wrappers. Filters [2][3][4] select the best subset of features using some predefined evaluation criteria that are independent of the learning algorithm, so, they are also called classifier-independent methods. On the other hand, wrappers are classifier-specific [5] in that they utilize the learning algorithm as the evaluation function and search for the best subset of features that optimizes the generalization performance. The wrapper methods may perform better, but involve huge computation. So, it is difficult for them to deal with high-dimension datasets.

Feature selection methods can also be grouped into supervised ones and unsupervised ones depending on whether class labels are considered or not. Supervised methods aim to choose the features that have the most discriminant information, measured by, e.g. mutual information [2], Fisher criterion [6] and Relief [7]. As a result, they remove noisy/irrelevant features which have little discriminant information. Unsupervised methods usually measure the similarity or correlation between features so as to remove the redundant ones [8]. Various similarity measures, like correlation coefficient and linear dependence, have been used. Recently, the combination of supervised and unsupervised feature selection [4] [9] by considering relevance and redundancy simultaneously has been reported. The measures of relevance and redundancy are mostly based on mutual information, which is difficult to calculate for continuous data except that special forms of probability density are assumed.

In this paper, we propose a new feature selection approach that combines the Fisher criterion and the principal feature analysis (PFA) technique [10]. PFA is an unsupervised technique that efficiently selects uncorrelated features, but does not consider the discriminant information of features. Using the Fisher criterion as a pre-selector to remove irrelevant features, PFA can then select a compact subset of relevant and uncorrelated features, which lead to high classification performance. We demonstrate the effectiveness of the composite approach in experiments on five publicly available datasets. The results show that our approach can largely reduce the feature dimensionality with little loss of classification accuracy.

The rest of this article is organized as follows: Section 2 describes the Fisher criterion for pre-selection; Section 3 describes the PFA technique for removing redundant features. Section 4 gives the flow of our composite algorithm of feature selection. Section 5 presents our experimental results, and Section 6 gives concluding remarks.

1-4244-0973-X/07/$25.00 ©2007 IEEE

2. Fisher Criterion for Pre-Selection

Consider the classification of C classes. Given n_i training samples (vectors) $\{\mathbf{x}_1^i, \mathbf{x}_2^i, \ldots, \mathbf{x}_{n_i}^i\}$ for each class i, ($i = 1, 2, \cdots, C$). The a priori probability of class i is estimated by $P_i = \frac{n_i}{\sum_{i=1}^{C} n_i}$. The class means μ_i are estimated by $\hat{\mu}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} \mathbf{x}_j^i$, and the gross mean μ is estimated by $\hat{\mu} = \sum_{i=1}^{C} p_i \hat{\mu}_i$.

The sample covariance matrix $\hat{S}_i$ of class i is estimated by

$$\hat{S}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} (\mathbf{x}_j^i - \hat{\mu}_i)(\mathbf{x}_j^i - \hat{\mu}_i)^T . \quad (1)$$

The within-class scatter matrix and between-class scatter matrix are estimated by

$$S_w = \sum_{i=1}^{C} P_i \hat{S}_i \quad (2)$$

and

$$S_b = \sum_{i=1}^{C} P_i (\hat{\mu}_i - \hat{\mu})(\hat{\mu}_i - \hat{\mu})^T , \quad (3)$$

respectively.

The Class Separability of a feature set can then be measured by

$$J_F = trace(S_w^{-1} S_b) . \quad (4)$$

This measure serves a good criterion for feature subset selection, and has shown superior performance in many practical problems. However, its calculation for a large number of features is computationally expensive. Instead, the Fisher criterion for one single feature has been prevalently used to select discriminant features. For the *k*th feature, it is calculated by

$$Fisher(k) = \frac{S_b^{(k)}}{S_w^{(k)}} , \quad (5)$$

where $S_b^{(k)}$ and $S_w^{(k)}$ are the *k*th diagonal element of S_b and S_w, respectively, and can also be calculated from the data of single feature.

For feature pre-selection, we calculate the Fisher criterion of each feature, order the features in decreasing order of criterion values, and simply select the features of maximum values, while the features with very small Fisher values are abandoned. Though the single-feature Fisher criterion does not consider the joint seperability of multiple features, it is able to retain all discriminant features by only removing irrelevant/noisy features, for which the Fisher criterion is nearly zero.

3. Principal Feature Analysis (PFA)

PFA [10] is a feature grouping (and consequently choosing a feature from each group) technique based on the popular principal component analysis (PCA) [11] technique. PCA is an unsupervised approach to project the original feature space onto a low-dimensional subspace, but obtains a set of transformed features rather than a subset of the original features. How to use the transformation matrix of PCA to select features has been reported in [10] [12].

3.1. Principal Component Analysis (PCA)

PCA projects *d*-dimensional vectors onto a subspace of lower dimensionality in the way that the square error of vector reconstruction is minimized [11]. To do this, the $d \times d$ covariance matrix is estimated on the training vectors, its eigenvectors and eigenvalues are calculated, and sorted in decreasing order of eigenvalues. Then, the *k* eigenvectors corresponding to the largest eigenvalues are used as the basis vectors of the subspace. A *d*-dimensional vector **x** is projected to *k*-dimensional subspace vector **y** by

$$\mathbf{y} = A^T(\mathbf{x} - \mu) , \quad (6)$$

Where μ is the mean of training vectors, and A is a $d \times k$ matrix composed by the principal eigenvectors as columns.

3.2. Principal Feature Analysis (PFA)

PFA exploits the structure of principal components to choose principal features, which retain most of the information both in the sense of maximum variability in low-dimensional subspace and in the sense of minimizing the reconstruction error [10][13].

Let X be an *n*-dimensional feature vector, $\sum$ is the covariance matrix of X. Let A be the orthonormal matrix composed of the eigenvectors of $\sum$:

$$\sum = A \Lambda A^T , \quad (7)$$

where $\Lambda = diag[\lambda_1, \lambda_2, \ldots, \lambda_n]$, with $\lambda_1 \geq \lambda_2 \geq \cdots \lambda_n$.

Let A_q be the matrix composed of the first q columns of A and $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n \in \mathrm{R}^q$ be the rows of A_q.

To find the representative feature subset, we use the row vectors $\mathbf{v}_i$ to cluster the features that are highly correlated and then choose one feature from each cluster.

The algorithm of PFA can be summarized in five steps:

1) Compute the sample covariance matrix.

2) Compute the principal components and eigenvalues of the covariance.

3) Choose the subspace dimension q and construct the matrix A_q from A. This can be chosen according to the desired variability of the data to be retained. The retained variability is the ratio between the sum of the first q eigenvalues and the sum of all eigenvalues. This ratio is empirically set to 90% in our experiments.

4) Cluster the vectors $|\mathbf{v}_1|, |\mathbf{v}_2|, \ldots, |\mathbf{v}_n| \in \mathrm{R}^q$ to p clusters using the k-means algorithm. $|\cdot|$ denotes the vector composed of the abstract values of all elements.

5) From each cluster, find the corresponding vector $\mathbf{v}_i$, which is closest in Euclidean distance to the mean of the cluster. Choose the corresponding feature, x_i, as a principal feature. This step will yield the choice of p principal features.

The key idea of PFA is that features which are highly correlated have similar absolute value of weight vectors $\mathbf{v}_i$ [10].

Like PCA, PFA is an unsupervised learning method, so, it is not sensitive to noise and irrelevant features. Unlike the PCA technique that generates transformed features (principal components), PFA selects a subset of preliminary features (principal features).

4. Composition of Fisher Criterion and PFA

The Fisher criterion is a supervised criterion. It can be used to remove the features which are noisy or irrelevant, but does not consider the redundancy of features. For example, if two feature vectors are entirely same and both have high Fisher values, they will be both selected with high redundancy. On the other hand, the method of principal feature analysis (PFA) is unsupervised. It explores the correlation between features and removes redundant ones, but cannot distinguish noisy features from relevant ones.

For removing both irrelevant and redundant features, we thus combine the Fisher criterion and PFA. The single-feature Fisher criterion is used as the pre-selection criterion to select the best m individual discriminant features. Then PFA is used to cluster the m pre-selected features into p groups, and one feature from each group forms a subset of principal features.

In pre-selection by Fisher criterion, how to decide the number of retained features is an issue. In our experiments we accumulate the Fisher values in decreasing order until the sum of values exceeds a pre-specified percentage of total values.

The composite feature selection algorithm is summarized in four steps as follows.

Pre-selection

1) Calculate the Fisher criterion value for each feature in the preliminary set.

2) Order the features in decreasing order of Fisher value, select the leading features until the percentage of accumulated Fisher values exceeds 0.99. The m leading features are pre-selected ones.

Principal feature analysis

3) Using PFA to cluster the pre-selected features into p groups.

4) From each group, select the feature that is closest to the cluster center. The p principal features, one from each group, form the selected feature subset for classification.

The computation of the composite algorithm has three main parts: the computation of single-feature Fisher values, principal component analysis (PCA) in PFA, and clustering of weight vectors in PFA. The complexity of single-feature Fisher criterion is linear with the number n of preliminary features. PCA involves the orthogonal decomposition of $m \times m$ covariance matrix (m is the number of pre-selected features), and its complexity is $O(m^3)$. The complexity of k-means clustering is $O(m \times p)$ (p is the number of clusters). Hence, the computational complexity of the composite algorithm is mainly dependent on PCA. On the other hand, the PCA of preliminary feature set has complexity $O(n^3)$ (n>m). For the joint Fisher criterion on a feature set, the computation involves matrix decomposition and the complexity is $O(p^3)$. Using a sub-optimal sequential forward search algorithm for selection, the total complexity of joint Fisher criterion is $O(n \times p^4)$ (p is the number of selected features). Overall, our composite algorithm has a complexity comparable to PCA, and much lower than joint Fisher criterion J_F.

5. Experiments

To demonstrate the effectiveness of the proposed feature selection approach, we conducted experiments of pattern classification on five publicly available datasets.

5.1. Datasets

Table 1 gives a summary of the datasets used in our experiments. All the datasets except for USPS and Leukemia were obtained from the UCI Repository [14]. The USPS dataset contains handwritten digit images, which is available from [15]. The Leukemia dataset is a gene dataset which is available form [16].

For Multi-feat. dataset, we evaluate classification performance by 5-fold cross-validation. The other datasets are divided into standard training and test subsets. On each partition of a dataset, feature selection and classifier design are performed on the training subset, and classification accuracy is evaluated on the test/validation subset. In cross validation, the accuracy is averaged over 5 partitions.

The features in Multi-feat. dataset have significantly variable scales. We normalized the data such that each feature has zero mean and standard deviation 1. The other four datasets remain un-normalized.

In classification, we use two types of classifiers: nearest neighbor (1-NN) classifier and support vector machine (SVM) classifier with 4-th order polynomial kernel. The SVM classifier combines one-vs-all binary classifiers for multi-class classification.

For comparison with the accuracies after feature selection, the baseline accuracies on the preliminary feature set (containing all features) are given in Table 1.

5.2. Experimental results

In feature selection using the proposed composite approach, the number of pre-selected features is determined by accumulated Fisher values until 99% of total values. For Leukemia datasets with many preliminary features, the number of pre-selected features is set to 300.The number of features selected by PFA is set to be about half of the number of pre-selected features.

Table 2 provides the experimental results of pre-selection by Fisher criterion and final selection by principal feature analysis (PFA). For Multi-feat., since we select and evaluate features using 5-fold cross-validation, the number of features after pre-selection varies with training subset. We hence give for them the standard deviation of the number pre-selected features as well as the average number.

It is shown in Table 2 that after pre-selection, a few features with small values of Fisher criterion are deleted, and the accuracy is not decreased (compared to the baseline accuracies) considerably. For some datasets (Isolet, Leukemia), the accuracies even increase after pre-selection of features. That is because there are some noisy/irrelevant features that contain little discriminant information in the preliminary feature space. The removal of them thus improves the generalization performance.

Table 1. Summary of experimented datasets

Dataset	#Feature	#Class	#Sample	Baseline accuracy (%)	
			training/test	1NN	SVM
Musk	166	2	6598/476	97.48	98.11
USPS	256	10	7291/2007	94.77	96.01
Isolet	617	26	6238/1559	88.58	96.73
Multi-feat.	649	10	2000	98.20	98.80
Leukemia	12558	7	215/112	89.29	92.96

Table 2. Classification accuracies after feature selection

Dataset	#Feature pre-selected	Accuracy after Pre-selection (%)		# Feature after PFA	Accuracy after PFA (%)	
		1NN	SVM		1NN	SVM
Musk	129	96.01	97.06	65	97.06	97.06
USPS	234	94.87	95.86	117	94.32	95.47
Isolet	571	89.42	96.79	286	88.13	96.54
Multi-feat.	594(0.75)	98.10	98.75	297	97.95	98.85
Leukemia	300	90.18	94.64	150	90.18	92.86

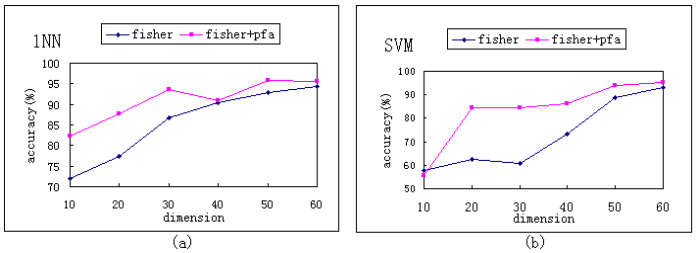

Figure 1. Test accuracy of Musk dataset. (a) 1NN classifier accuracy; (b) SVM classifier accuracy.

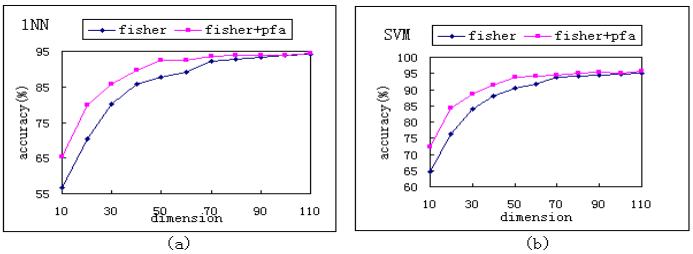

Figure 2. Test accuracy of USPS dataset. (a) 1NN classifier accuracy; (b) SVM classifier accuracy.

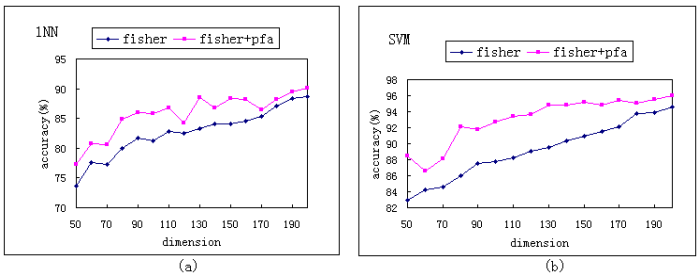

Figure 3. Test accuracy of Isolet dataset. (a) 1NN classifier accuracy; (b)SVM classifier accuracy.

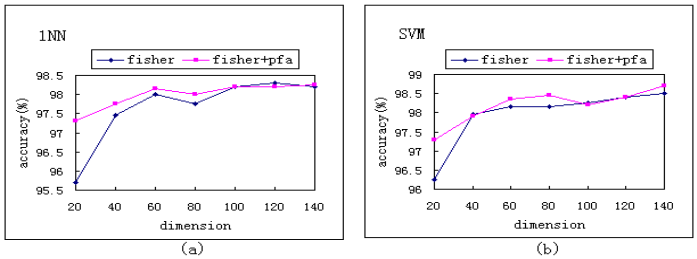

Figure 4. Cross-validation accuracy of Multi-feat. dataset. (a) 1NN classifier accuracy; (b)SVM classifier accuracy

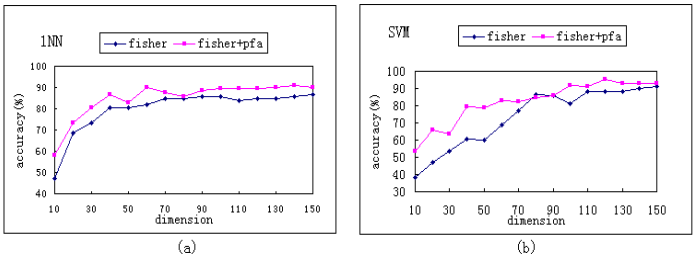

Figure 5. Test accuracy of Leukemia dataset. (a) 1NN classifier accuracy; (b)SVM classifier accuracy

Comparing the classification accuracies of feature selection by PFA with those by pre-selection (Table 2), we can see that the number of features is reduced remarkably with the accuracy decreased only slightly. This indicates that these datasets have many redundant features, which can be effectively removed by PFA.

On the other hand, for proving that PFA can remove the redundant features, we design another experiment. First, we select the best *2d* dimension feature by using Fisher criterion, and then we use PFA to reduce the dimension to *d* dimension. We compare this classification accuracy with selecting the best *d* dimension feature by using Fisher criterion. Figure 1-5 show the experimental results on the five datasets.

6. Conclusion

In this paper we proposed a composition method for feature selection. The method is combining Fisher criterion and principal feature analysis (PFA). Because PFA is an unsupervised method, we use Fisher criterion to delete the features that contain less discriminating information. The PFA chooses the principal features, which contain most of the information. The proposed method is applied to some datasets, the results shows that the composition of Fisher criterion and PFA is an available method for feature selection.

Acknowledgements

This work was performed at Institute of Automation, Chinese Academy of Sciences (CAS), and was supported by the Hundred Talents Program of CAS and the Natural Science Foundation of China (NSFC) under the grant No.60121302.

References

[1] John. G.H., Kohavi. R., and Pfleger. K., "Irrelevant feature and the subset selection problem", Proceeding of the 11th International Conference on Machine Learning, San Francisco, pp.121-129, 1994.

[2] Kwak. N., and Choi. C. H., "Input feature selection by mutual information based on Parzen window", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 24, No. 12, pp.1667-1671, 2002.

[3] Hall. M., "Correlation-based feature selection for discrete and numeric class machine learning", Proceeding of the 17th International Conference on Machine Learning, Stanford, pp. 259-266, 2000

[4] Yu. L., and Liu. H., "Efficient feature selection via analysis of relevance and redundancy", Journal of Machine Learning Research, Vol. 5, pp. 1205-1224, 2004

[5] Kudo. M., and Sklansky. J., "Comparison of algorithms that select features for pattern classifiers", Pattern Recognition, Vol. 33, No. 1, pp. 25-41, 2000.

[6] Duda. R. O., Hart. P. E., and Stork. D. G., Pattern recognition (2nd edition), John Wiley & Sons, California, 2000

[7] Kononenko. I., "Estimating attributes: analysis and extension of RELIEF", Proceeding of European Conference on Machine Learning, Berlin, pp. 171-182, 1994

[8] Mitra. P., Murthy. C.A., and Pal. S. K., "Unsupervised feature selection using feature similarity", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 24, No. 3, pp. 301-312, 2002.

[9] Peng. H., Long. F., and Ding. C., "Feature selection based on mutual information criterion of max-dependency, max-relevance, and min-redundancy", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 27, No. 8, pp.1226-1238, 2005.

[10] Cohen. I., Tian. Q., Zhou. X.S., and Huang. T.S., "Feature selection using principal feature analysis", Beckman Institute for advanced science and technology university of Illinois , 2002

[11] Jolliffe. I.T., Principal Component Analysis, Springer-Verlag, New-York,1986

[12] Krzanowshi. W.J., "Selection of variables to preserve multivariate data structure using principal component analysis", Applied Statistics, Vol.36, No.1, pp.22-23, 1987

[13] Yoon. H., Yang. K., and Shahabi. C., "Feature subset selection and feature ranking for multivariate time series", IEEE Transactions on knowledge and data engineering, Vol. 17, No. 9, pp.1186-1198, 2005

[14] Merz. C.J., and Murphy. P.M., UCI repository of machine learning databases, Univ. of California Irvine, http://www.ics.uci.edu/~mlearn/MLRepository.html

[15] http:// www.kernel.org/data.html

[16] http://www.stjuderesearch.org/data/all1/