A method for feature selection based on the correlation analysis

Jinjie Huang, Ningning Huang, Luo Zhang, Hongmei Xu
Department of Automation
Harbin University of Science and Technology
Harbin, China
jjhuang@hrbust.edu.cn, anhnn@163.com, zhangluo1990@sina.com

Abstract—Feature selection is one of the important issues in the fields of machine learning and pattern classification. The classification ability of features is analyzed from the point of view of correlation and redundancy. Two types of correlation: C-correlation and F-correlation are presented. The C-correlation is applied to identify the relevant features to the category attribute, while the F-correlation is used to measure the redundancy among features. Finally, the dimension of input features is further reduced with the sequential forward search strategy. Thus a method for feature selection based on the correlation analysis of features is derived. The experimental results show that the proposed algorithm is an effective method for feature selection.

Keywords- feature selection; correlation; redundancy; dimension reduction

I. INTRODUCTION

Feature selection is one of the most important issues in the fields of machine learning and pattern recognition. It aims to choose some most effective features to reduce the dimension of the feature space [1, 8]. Many huge data sets of high dimensions usually contain a lot of redundant even uncorrelated features. This will result in a low efficiency to train the learning algorithm and much computational complexity, or even some loss of the classification accuracy. We can remove the uncorrelated and redundant features from high dimensional data set to choose an optimal subset of features. This method can increase the efficiency of learning algorithm and reduce the computational complexity effectively [1,2].

Feature selection based on correlation analysis is to investigate the size of the correlation between features and remove the related features to ensure the size of correlation to be small as far as possible. This method could achieve good classification results using as few features as possible.

Theoretically speaking, the most reliable search method for the optimal subset of features is the exhaustive method. But it is difficult to realize because of the large computational complexity caused by large dimension of feature space. Many scholars come up with plenty of effective search strategies which need to be integrated with different evaluation criteria. For the time being, there are several feature evaluation criteria such as distance measure, information measure, consistency measure, classification error rate. Yet, it is always a hot topic how to combine these criteria with the search strategies effectively to reveal the correlation and redundancy of candidate feature sets.

In this paper, two concepts of C-correlation and F-correlation are presented. With the two definitions we remove the features independent of the classification target at first, and then delete the redundant features correlated highly with other features. After that, we perform the searching operations in the subspace consisted of the features left. Finally, a feature subset, which is of little correlation mutually and with good classification ability, is obtained. Thus, the dimension of the original feature space is reduced, and the computational complexity for training a classifier is cut down greatly.

II. THE CORRELATION AND REDUNDANCY OF FEATURES

Correlation between input features and category attribute can be divided into three cases: strong correlation, weak correlation, and no correlation. A good input feature for classification should be strongly correlated with category attribute and weekly or no correlated with other input features, that is to say, it is not redundant. Features redundancy is referred to as the correlation among features. If two features are completely correlated, they are referred to as redundant features. As a good feature subset, its features should be strongly correlated with categories but uncorrelated with each other.

Correlation measurement can be generally divided into two categories: one is linear correlation such as linear correlation coefficient, pearson product moment correlation. The other one is based on information theory, such as entropy. Here, we use correlation coefficient r as the measurement of correlation for features.

Correlation efficient formula:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y} \qquad (1)$$

Where, x, y are two features, $\bar{x}$ 、 $\bar{y}$ are means of feature x and y:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \tag{2}$$

s_x and s_y are standard deviations of feature x and y:

$$s_x = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}, \quad s_y = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}} \tag{3}$$

Where n is the number of samples. r_{xy} ranges from -1 to 1. The larger the absolute value of r_{xy} is, the more strongly x and y are correlated, and that is to say, the greater the redundancy of x and y is. If the value of r_{xy} is zero, that means the two variables are independent each other.

Let's suppose that the feature set is $S = [F_1, F_2, ..., F_N]$, N is the total number of features, C is category attribute. Here, we define two kinds of correlations for input features.

Definition 1: For any feature $F_i \in S$, the correlation of F_i with category C is called C-correlation of feature F_i.

Definition 2: For any feature $F_i \in S$, the correlation of F_i with feature F_j ($j \neq i$) is called F-correlation of feature F_i.

III. Feature Selection Based on Correlation Analysis

A. C-correlation analysis

As mentioned above, the strength of correlation between a feature and the classification label determines the classification performance. The stronger the correlation is, the better the classification performance is. Therefore, firstly we should remove the features uncorrelated with class attribute in order to find an optimal feature subset. Here, we use the C-correlation of a feature as a measure to reduce the original feature dimensions by choosing all strongly correlated features and some weakly correlated features.

In order to improve the efficiency of the algorithm, we predefine a threshold value g ($g > 0$). Calculate the C-correlation of each future. If the value of C-correlation of a feature F_i isn't less than the predefined threshold g, i.e. $|r(F_i, C)| \geq g$, that means this feature has a considerable correlation with category C. Thus we can identify feature F_i as the one correlated with category. The selected features will be used for redundancy analysis further.

B. F-correlation analysis

In an ideal feature set, every feature should be correlated with class labels and uncorrelated each other. Therefore, we should remove redundant features after c-correlation analysis.

We still adopt the correlation coefficient as a measure for redundancy among features. Firstly, rank the features according to their C-correlation values; then calculate the F-correlation of each feature with others, and arrange their absolute values as the following matrix R:

$$R = \begin{pmatrix} 1 & r_{12} & r_{13} & \cdots & r_{1n} \\ & 1 & r_{23} & \cdots & r_{2n} \\ & & \ddots & r_{ij} & \vdots \\ & & & 1 & r_{(n-1)n} \\ & & & & 1 \end{pmatrix} \tag{4}$$

r_{ij} is the absolute value of F-correlation coefficient between feature i and j. Predefine a threshold δ , $(\delta > 0)$. When $r_{ij} > \delta$, remove the posterior feature and then keep on investigating the remaining features.

C. Optimal feature subset selection

For high dimensional data, the feature space is sometime still very large after removing the uncorrelated and redundant features. In order to reduce the computational complexity of classification, we optimize the remaining feature subset to reduce the dimension further.

Here we adopt sequential forward selection method to select features and use the correlation coefficients r for feature evaluation as usual. First, we select two features with minimum correlation coefficient, and then choose the next feature which has minimum correlation coefficient with the selected features. Continue selecting features from the rest until the stopping condition is satisfied. The structure of the algorithm is as Figure 1.

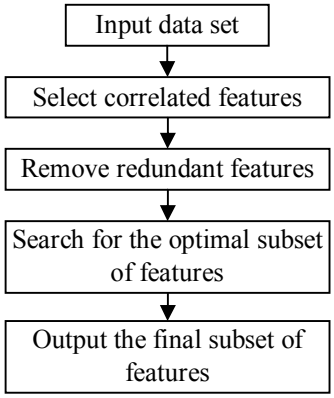

Figure 1. The structure of the algorithm

So far, the algorithm is as follows:

(1) Select correlated features. Calculate the C-correlation value of each feature with category. Preset a threshold g ($g > 0$). Select the features whose C-correlation is bigger than g.

(2) Remove redundant features. Rank the features from big to small by their C-correlation values. Calculate the F-correlation coefficient matrix R of features. Predefine a threshold δ , $(\delta > 0)$. When F-correlation coefficient is bigger than δ , remove the posterior feature. When one of the two

Identify applicable sponsor/s here. If no sponsors, delete this text box. (sponsors)

features is removed, keep the rest one in the selected subset of features.

(3) Optimize the subset of selected features. Adopt sequential forward selection strategy. First, choose two features with minimum correlation coefficient, then choose the next feature which has minimal correlation coefficient with selected features from the remaining. The search ends when the classification accuracy is improved no longer or little.

IV. EXPERIMENTAL RESULTS

Experiments are performed on four data sets. The first one is *USPS* data set, which is of 10 handwritten numerals. The other three are from UCI data sets. Here, *waveform*40 is an artificial data set. It derives from the data set *waveform*21 by adding 19 dimensional noise features. The four experimental data sets are shown in Table 1.

The classifier used in the experiments is 5-nearest Neighbor (5NN). For every data set, we calculate the C-correlation, the F-correlation, and search the optimal feature subset step by step, record the size of feature subsets and the classification accuracies on test sets. In order to reserve the useful information as far as possible and avoid the loss of important features for classification, we usually take a small value for the threshold in C-correlation, here g=0.015. If the F-correlation coefficient value of two features is between 0.8-1.0, we can consider the two features are highly correlated. So we take δ =0.89. At last, in order to reduce the dimensions of feature space further, we carry on the searching operation in the selected feature subsets until the least number of features is found for different data sets.

TABLE I. . EXPERIMENTAL DATA SETS

	data sets	dimensions	test/training samples	category num.
1	waveform40	40	1000 /4000	3
2	isolet	617	1559/6238	26
3	mfeat	649	500 /1500	10
4	USPS	256	2007/7291	10

Just for comparison purposes, the size and classification accuracy of both original feature sets and feature subsets S1and S2, which are selected after the C-correlation and F-correlation stages, are listed in Table 2. The number of features selected and classification accuracy after searching operation in the third stage are also listed in Table 2. According to the tables, we can see that the dimension of feature space is greatly reduced after the three stages, but the ability of classification hasn't significantly decreased in low dimensional space, even improved a bit in *mfeat* and USPS data sets.

In addition, we can find that in most cases, the classification accuracy on feature subset S1 is higher than that on the original feature set S to a certain extent. This demonstrates that deleting uncorrelated features can improve the performance of learning algorithm. For data set *waveform*40 and UPS, the dimension of feature space S2 hasn't reduced anymore compared with that of S1 after C-correlation and F-correlation. This means that features in S1 are of low redundancy. For data sets *isolet* and *mfeat*, after F-correlation and optimal feature subset searching, the dimensions decrease sharply. This means that the redundancy of features in the two data sets is relatively large and deleting these features has no effect on the performance of learning algorithm. For data sets *waveform*40, the dimensions have been relatively low and the classification accuracy remains unchanged after C-correlation and F-correlation, but after optimal subset searching operation, the classification accuracy decreases obviously. This illustrates that the data set S2 has approached optimal and doesn't need further selection. The accuracy of *isolet*, *mfeat*, and *USPS* has little change but the dimension is reduced considerably after optimal subset searching operation. This illustrates that it is necessary to adopt dimension reduction process after removing uncorrelated and redundant features for high dimensional data sets.

TABLE II. DIMENSION OF FEATURE SPACE IN DIFFERENT STAGES AND CLASSIFICATION ACCURACY

Data sets	Original dim.	Acc. (%)	Dim. after C-correlation	Acc.(%)	Dim. after F-correlation	Acc.(%)	Dim. after searching	Acc. (%)
waveform40	40	80.10	19	81.70	19	81.70	15	79.4
isolet	617	91.98	565	91.60	402	91.02	297	90.64
mfeat	649	96.00	607	96.00	487	97.40	358	97.20
USPS	256	94.42	246	96.66	242	96.71	110	97.16

V. CONCLUSIONS

Through study on two types of correlation: C-correlation and F-correlation, we analysis the classification ability of feature subsets obtained by the C-correlation and F-correlation operations. On this basis, we propose a method for dimensional reduction with correlation analysis. The method is simple and rapid. It can preserve the original information of features at their maximum and has good classification efficiency and applicability. However, how to determine the number of features in the optimal feature subset appropriately is still the research work in future.

ACKNOWLEDGMENT

This work is partially supported by the Science Foundation of Educational Department of Heilongjiang Province of China under Grant No.12511105 and the Science and Technology Foundation for Innovative Talents of Harbin City of China under Grant No. 2007RFXXG023.

REFERENCES

[1] LANGLY P. Selection of relevant features in machine learning [C]// Proceedings of the AAAI Fall Symposium on Relevance. New Orleans, LA: AAAI PRESS, 1994:140-144

[2] JOHN G,KOHAVIR, PFLEGER K. Irrelevant feature and the subset selection problem[C]//Proceedings of the 11th International Conference on Machine Learning. New Brunswick, NJ, USA : Morgan Kaufman Publishers,1994:121-129.

[3] JI Xiao-jun, LI Shi-zhong, LI Ting. Application of the Correlation Analysis in Feature Selection[J]// JOURNAL OF TEST AND MEASUREMENT TECHNOLOGY, 2001,15(1):15

[4] DUDARO, HART P E, STORK D G. Pattern classification [M]. 2nd ed. New York: John WIELY&Sons, 2001.

[5] MERZ C J, MURPHY PM. UGI repository of machine learning data bases[DB/OL]. [2007-05-01]. http://www.ics.uci.edu~,mlearn/MLRepositort.html

[6] HE ZHI-WEN,etc. A study on methods of feature selection based on the correlation analysis[J]. Nuclear Electronics & Detection Technology,2005,11(6):25.

[7] LI YUN, YE CHUN XIAO. Study on Feature selection based on feature correlation, 2004,6.

[8] Isabelle Guyon , NIPS 2001 workshop on variable and Feature Selection : 2001, December 6-8, [http: //www.Clooinet.com/isabelle/Proiects/NIPS2001/].

[9] Anil K. Jain, Robert P. W. Duin, Jianchang Mao. Statistical Pattern Recognition : A Review[J]. IEEE Transactions on Patern Analysis and Machine Intelligence, 2000, 22(1) : 4-37

[10] M. Dash and H. Liu . Feature Selection for Classification[J]. Intelligent Data Analysis, 1997, 1(3) : 131-156.