An Improved Feature Selection Algorithm Based on Parzen Window and Conditional Mutual Information

He Deng-chao[1], Hao Wen-ning[1], Chen Gang[1], Jin Da-wei[1]
Command Information Systems Institute
PLA University of Science &Technology
Nanjing 210007 China
Email:tkod@qq.com

***Abstract*—In this paper, an improved feature selection algorithm by conditional mutual information with Parzen window was proposed, which adopted conditional mutual information as an evaluation criterion of feature selection in order to overcome the deficiency of feature redundant and used Parzen window to estimate the probability density functions and calculate the conditional mutual information of continuous variables, in such a way as to achieve feature selection for continuous data.**

Keyword-Feature Selection; Conditional Mutual Information; Parzen window

I.INTRODUCTION

Feature selection has become an indispensable part of data preprocessing in many fields(e.g. text classification, image search, biological information processing) especially high dimensional data preprocessing. This feature selection algorithm which uses an evaluation criterion to select relevant features from a large dataset is performed in order to reduce hypothesis space of search and storage, and enhance the performance of the data mining. The existing approaches of feature selection can be divided into two categories: Wrapper and Filter[1]. As Wrapper method, the performance of a particular inductive learning algorithm would be adopted as criterion of feature evaluation and selection, such as feature gene subset algorithm based on classification error and selection algorithm based on SVM etc. The deficiency of Wrapper is that inheriting the offset of used inductive learning algorithm, thus this method only has good performance in preselected inductive learning algorithm, so it has no generality. In addition, because of high computational complexity, it may not be applicable to big dataset. Therefore when dataset has a large number of features, Filter method has better performance than Wrapper. Filter method is a feature selection technique that adopting a specific evaluation criterion to select features, which is independent of inductive learning algorithms. The conventional evaluation criterions of Filter method adopted are: χ^2-test [2], information entropy[14], mutual information[4], minimum joint mutual information loss[15], minimum classification error[6], etc.

Information theory based evaluation criterion could well reflect relevance between two random variables so that it's usually adopted in feature selection problem. [15] used conditional mutual information as criterion of feature evaluation, which considered relevance not only between feature and class but also within features. This method has good performance only for discrete variables, however， for continuous variables, data should be discretized before relevant calculation. Data discretization partitions would result the lost of original data information. For a same continuous variable, with the decrease of the number of discretization partitions, absolute value of Kendall coefficient between variable and dependent variable would become larger and absolute value of Spearman coefficient would become smaller. Therefore, it can be seen that the solution of data discretion could affect the result of relevant calculation. Another solution is that processing under the assumption that probability distribution of data is already known[11]. But in fact the assumed distributions seldom meet the practical situation. [9] has proposed a mutual information and Parzen window based feature selection method which adopted mutual information as evaluation criterion and used Parzen window to estimate probability density functions(pdf) of continuous variable in order to compute mutual information between different continuous variables. However, this method only considered relevance between single feature and class label variable and did not consider relevance within features, thus it may result feature redundant. From the above, for existing information theory based feature selection methods, there are two main deficiencies about relevance calculation:

1, lack of effective relevant calculation method for continuous variables;

2, difficult to calculate relevance both between feature and class and within features comprehensively;

According to the above deficiencies, an improved feature selection algorithm was proposed in this paper, which adopted conditional mutual information as an evaluation criterion of feature selection in order to overcome the deficiency of feature redundant and used Parzen window to estimate the probability density functions and calculate the conditional mutual information of continuous variables, in such a way as to achieve feature selection for continuous data.

II.PRELIMINARIES

A.Information Theory

In information theory, the entropy is a measure of

Published by Atlantis Press, Paris, France.

random variable. The entropy of X is defined as[8]

$$H(X) = -\sum_{x\in X} p(x)\log_2 p(x) \quad (1)$$

Given a certain variable X, the uncertainty of the other correlated variable Y can be measured by the conditional entropy[8]

$$H(Y|X) = -\sum_{x\in X}\sum_{y\in Y} p(x,y)\log_2 p(y|x) \quad (2)$$

Mutual information is introduced to measure relevance between two random variables[8]

$$I(X;Y) = \sum_{x\in X}\sum_{y\in Y} p(x,y)\log_2 \frac{p(x,y)}{p(x)p(y)} \quad (3)$$

The mutual information between feature X and class C could be defined as[8]

$$I(X;C) = H(C) - H(C|X) \quad (4)$$

For continuous random variables, the entropy and mutual information are defined as

$$H(X) = -\int p(x)\log_2 p(x)dx \quad (5)$$

$$I(X;Y) = \int p(x,y)\log_2 \frac{p(x,y)}{p(x)p(y)} dxdy \quad (6)$$

If S is a known feature set, the conditional mutual information $I(C;f_i|S)$ between class C and feature f_i, is defined as[11]

$$I(C;f_i|S) = I(C;S,f_i) - I(C;S) \quad (7)$$

The conditional mutual information could be calculated by mutual information in (7).

B.The Parzen Window Density Estimate

Parzen window density estimate is an non-parametric estimation method with sound theoretical foundations and excellent performance which could approximate the probability density of a continuous variable very well. Its basic idea is to estimate the overall density function by mean value of the density of every points in a certain domain. In generally, supposed that x is a point in d-dimensional space and N is the size of sample, in order to estimate distribution probability density $p(x)$, a hypercube V with length h and centre point x is constructed, and its volume $V = h^d$. To calculate the amount of sample N_V in V, a function is constructed as

$$\phi(u) = \begin{cases} 1, |u_i| \le \frac{1}{2}, i = 1,2,\cdots,d \\ 0, others \end{cases} \quad (8)$$

$\phi(u)$ satisfy $\phi(u) \ge 0$ and $\int \phi(u)du = 1$. The amount of sample in volume V is given by $N_v = \sum_{i=1}^{N} \phi(\frac{x-x_i}{h})$, then the probability density estimate here is given by

$$\hat{p}(x) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{V}\phi(\frac{x-x_i}{h}) \quad (9)$$

(9) is the basic formula of Parzen window density estimate[12] where $\phi(u)$ is called window function.

III.AN IMPROVED FEATURE SELECTION METHOD

A.Feature Selection by Conditional Mutual Information with Parzen Window

Because classification learning algorithm would construct a mapping rule for classification from a given training data set, there would be a mapping rule from feature set F to class C [5] for each class of each sample in the sample set. Hence, if a feature has important information which could influence class distribution, the feature is a relevant feature, otherwise the feature is an irrelevant feature or a redundancy feature. Since mutual information is considered as a good indicator of relevance between variables, it would be used to be the basic of valuation criterion which measures the relevance between feature set and class.

Given a feature set S ,if $I(C;f|S) > 0$, f is a relevant feature which has information about class C that the feature set S do not have. If feature f and class C are independent, i.e. $I(C;f|S) = 0$, f does not have any useful information for classification when S has been given, i.e. f is an irrelevant feature or a redundancy feature when S has been given. The conditional mutual information $I(C;f|S)$ measures relevance not only between feature and class but also between features(i.e. redundancy judgment), therefore it is appropriate to adopt conditional mutual information as evaluation criterion in feature selection.

Therefore the evaluation criterion of feature selection is given by

$$J(f_i) = I(C;f_i|S) \quad (10)$$

From (10), if the evaluation criterion $J(f_i)$ is large, it means feature f_i and class C are closely related, i.e. the possibility of feature f_i being selected is large.

The improved feature selection algorithm is realized as follows:

Input：A training dataset $U(F,C)$

Output：Selected feature subset S

1. Initialize：$S = \varnothing$.

2. $\forall f_i \in F$,compute $J(f_i)$.

3. (Selection of the first feature) find the feature that maximizes $J(f_i)$,set $F = F - \{f_i\}, S = S + \{f_i\}$.

4. repeat until desired number of features are selected.

a. $\forall f_i \in F$,compute $J(f_i)$

b. (Selection of the next feature) choose the feature $f_i \in F$ that maximizes $J(f_i)$, and set

$F = F - \{f_i\}, S = S + \{f_i\}$.

5. Output the set S containing the selected features.

For discrete data, the algorithm could easily compute the evaluation criterion $J(f_i)$. However, for continuous data, to compute the conditional mutual information, probability density function must be known, but this is difficult in practice. To avoid this practical obstacle, a solution that data discretization before relevance calculation is normally adopted, which would result lost of original data information; another solution is that processing under the assumption which assumes probability distribution of data is already known[11] , but in fact the assumed distributions seldom be accordant with the practical situation. To overcome these problems, this paper proposes a new method for computing the conditional mutual information in the following subsection.

B.Calculation of Conditional Mutual Information with Parzen Window

In classification problems, the class normally has discrete values while the feature variables have continuous values. In this case, the conditional mutual information could be calculated by mutual information in (7). Then in (4), because the class is a discrete variable ,the entropy of the class variable $H(C)$ can be easily achieved in (1).But the conditional entropy[9]

$$H(C|X) = -\int_X p(x) \sum_{c=1}^{N} p(c|x) \log_2 p(c|x) dx \quad （11）$$

Where N is the number of class, since $p(c|x)$ is difficult to estimate, (11) is not easy to compute.

To overcome this difficulty, this paper proposes a method of calculating the conditional entropy and the conditional mutual information with Parzen window probability density estimate. By the Bayesian rule, the conditional probability can be written as

$$p(c|x) = \frac{p(x|c)p(c)}{p(x)} \quad （12）$$

For a certain feature, since its probability density distribution is unknown, the key of calculation is that how to estimate the probability density distribution of continuous variables, i.e. how to get $p(x|c)$. In this paper, Parzen window is adopted to estimate the probability density. In basic formula of Parzen window, window width h is an important parameter which has great influence to the result of estimate when the number of sample N is limited. The probability density of data of a given feature in each class can be estimated as[9]

$$\hat{p}(x|\ c) = \frac{1}{n_c} \sum_{i \in I_c} \phi(x - x_i, h) \quad （13）$$

Where I_C is the number of the training sample belonging to class C. Because the summation of the conditional probability equals one, i.e., $\sum_{k=1}^{N} p(k|x) = 1$,the conditional probability $p(c|x)$ is written as[9]

$$p(c|x) = \frac{p(c|x)}{\sum_{k=1}^{N} p(k|x)} = \frac{p(c)p(x|c)}{\sum_{k=1}^{N} p(k)p(x|k)} \quad （14）$$

The second equality is by (11). Then taking the estimate formula of $p(x|c)$ (16) into (14), the estimate formula[9] of the conditional probability $p(c|x)$ is given by

$$\hat{p}(c|\ x) = \frac{\sum_{i \in I_C} \phi(x - x_i, h_c)}{\sum_{k=1}^{N} \sum_{i \in I_k} \phi(x - x_i, h_k)} \quad （15）$$

Where h_c , h_k are window width parameters.

If Gaussian window function has been selected, (17) [9] would be rewritten as

$$\hat{p}(c|\ x) = \frac{\sum_{i \in I_C} \exp(-\frac{(x - x_i)^T \sum^{-1} (x - x_i)}{2h^2})}{\sum_{k=1}^{N} \sum_{i \in I_k} \exp(-\frac{(x - x_i)^T \sum^{-1} (x - x_i)}{2h^2})} \quad （16）$$

Then taking (16) into (11), the estimate formula[9] of conditional entropy is achieved by

$$\hat{H}(C|X) = -\sum_{j=1}^{n} \frac{1}{n} \sum_{c=1}^{N} \hat{p}(c|x_j) \log_2 \hat{p}(c|x_j) \quad （17）$$

From the above, the conditional mutual information for continuous variables can be easily and conveniently calculated by the proposed Parzen window based method.

IV.EXPERIMENTS AND RESULTS ANALYSIS

In this section, this paper apply the proposed improved feature selection algorithm to some of the specific classification problems and show the effectiveness of the proposed method through comparing experiment results. All the following experiments are performed on windows XP SP3, the feature selection algorithm is realized on Matlab R2008a. For convenience, the proposed method is referred as PWIFS(Parzen window improved feature selector) from now on.

A.Sonar Dataset

Sonar dataset was used in [3], [5], [9] to test the performances of their feature selection method. It consists of 208 patterns and has 60 features and two classes: metal and rock. The number of patterns belonging to metal is 111,and other 97 patterns are belonging to rock. Using PWIFS to select features of the dataset, and selecting 2~10 features among the 60 features to classification experiments and results analysis. To avoid overtraining, 50% of the dataset is used as training set and the other 50% is used as test set. Experiment results are showed in Table 1, the performance of PWIFS is compared with PWFS, MIFS, MIFS-U. In the table, all the resulting classification rates are the average value of 10 experiments and the corresponding standard deviations are shown in the parentheses.

TABLE 1 CLASSIFICATION RATES WITH DIFFERENT NUMBERS OF FEATURES FOR SONAR DATASET(%)
(NUMBER IN THE PARENTHESES ARE THE STANDARD DEVIATIONS OF 10 EXPERIMENTS)

Number of feature	PWIFS	PWFS	MIFS	MIFS-U
2	75.5（1.5）	71.8（2.1）	51.7（2.1）	65.2（1.6）
4	76.6（1.1）	76.6（3.1）	74.8（1.4）	77.3（0.4）
6	78.4（1.5）	78.4（1.5）	76.5（2.4）	77.9（0.7）
8	78.5（0.5）	78.5（0.5）	77.2（3.1）	78.9（0.8）
10	80.9（1.9）	80.9（1.9）	78.1（1.8）	81.5（0.4）
ALL（60）	87.9（0.2）			

From the table, in can be seen that PWIFS produced better performances than the others.

B.Other UCI Datasets

PWIFS was tested for various datasets in the UC-Irvine repository and compared the performance with other feature selection algorithm. Table 2 is the brief information of the datasets used in this paper.

TABLE 2 BRIEF INFORMATON OF THE DATASETS USED

Name	#features	#instances	#classes
Letter	16	20000	26
Breast Cancer	9	699	2
Waveform	21	1000	3
Vehicle	18	946	4

For these datasets, 4 relevant features are selected to classification experiments, and the results are shown in Table 3. In the experiments, , 75% of each dataset are used as training set and the other 25% as test set. From the table, it can be seen that PWIFS obtains better performances than other algorithms.

TABLE 3 CLASSIFICATION RATES(%) FOR UCI DATASETS

Datasets	PWIFS	PWFS	MIFS	MIFS-U
Letter	69.1	67.5	62.4	68.5
Breast Cancer	96.8	96.6	93.7	94.2
Waveform	76.5	75.4	67.6	73.8
Vehicle	63.8	62.5	57.3	59.9

V.CONCLUSIONS

In this paper, an improved feature selection algorithm was proposed, which adopted conditional mutual information as an evaluation criteria of feature selection and used Parzen window to estimate the probability density functions and calculate the conditional mutual information of continuous variables, in such a way as to achieve feature selection. Then PWIFS was applied for five datasets in UC-Irvine repository, from the results of experiments, it can be seen that PWIFS can select features with good performance, and for continuous variables, PWIFS exhibits better performance than the conventional methods such as PWFS, MIFS, and MIFS-U.

REFERENCES

[1] Guyon Ⅰ, Elisseeff A. An introduction to variable and feature selection[J].Journal of Machine Learning Research, 2003, 3:1157-1182

[2] Qu G, Hariri S, Yousif M. A new dependency and correlation analysis for features[J]. IEEE Transactions on Knowledge and Data Engineering, 2005, 17(9):1199-1207

[3] R. Battiti, "Using Mutual Information for Selecting Features in Supervised Neural Net Learning," IEEE Trans. Neural Networks, vol. 5, no. 4, pp. 537-550, July 1994.

[4] Liu H, Sun J, Liu L, et al. Feature selection with dynamic mutual information[J].Pattern Recgonition, 2009, 42(7):1330-1339

[5] N. Kwak and C.-H. Choi, "Input Feature Selection for Classification Problems," IEEE Trans. Neural Networks, vol. 13, no. 1, pp. 143-159, Jan. 2002.

[6] Sotoca J, Pla F. Supervised feature selection by clustering using conditional mutual information based distances[J]. Pattern Recognition, 2010,43(6):2068-2081

[7] T.M. Cover and J.A. Thomas, Elements of Information Theory. John Wiley &Sons, 1991.

[8] Cover T, Thomas J. Elements of Information Theory[M].NewYork:Wiley,1991

[9] Kwak N, Choi C H. Input feature selection by mutual information based on Parzen window[J]. Transactions on Pattern Analysis and Machine Intelligence, 2002,24(12).

[10]Quinlan R. C4.5:Programs for Machine Learning[M]. SanFrancisco: Morgan Kaufmann,1993

[11] BIAN Zhao-qi, ZHANG Xue-gong. Pattern Recognition, Second Edition. [M] .Beijing: Tsinghua University Press,2001.

[12] Richard O.Duba, Peter E.Hart, David G.Stork. Pattern Classification, Second Edition [M].John Wiley & Sons, Inc, 2001.

[13] Jiawei Han, Micheline Kamber. Data Mining: Concepts and Techniques, Second Edition [M]. Elsevier Inc, 2006.

[14] SONG Guo-Jie, TANG Shi-We, YANG Dong-Qing, WANG Teng-Jiao. A Spatial Feature Selection Method Based on Maximum Entropy Theory [J]. Journal of Software, 2003, 14(9)：1544-1550

[15] HUANG Jin-Jie, LV Ning, LI Shuang-Quan, CAI Yun-ze. Feature selection for classification analysis based on information-theoreric criteria[J]. ACTA Automatica Sinica, 2008,34(3)