Bio-Medical Materials and Engineering 26 (2015) S1953–S1959
DOI 10.3233/BME-151498
IOS Press

Feature gene selection method based on logistic and correlation information entropy

Jiucheng Xu[a,b,*], Tao Li[a] and Lin Sun[a,b]
[a]*College of Computer and Information Engineering, Henan Normal University, Xinxiang, China*
[b]*Engineering Technology Research Center for Computing Intelligence and Data Mining, Henan Province, China*

Abstract. In view of the characteristics of high dimension, small samples, nonlinearity and numeric type in the gene expression profile data, the logistic and the correlation information entropy are introduced into the feature gene selection. At first, the gene variable is screened preliminarily by logistic regression to obtain the genes that have a greater impact on the classification; then, the candidate features set is generated by deleting the unrelated features using Relief algorithm. On the basis of this, delete redundant features by using the correlation information entropy; finally, the feature gene subset is classified by using the classifier of support vector machine (SVM). Experimental results show that the proposed method can obtain smaller subset of genes and achieve higher recognition rate.

Keywords: Gene chips, logistic, correlation information entropy, feature selection

1. Introduction

With the development of gene expression profiles, the analysising and modeling of gene expression profiles have become an important topic in the field of bioinformatics research [1–3]. However, the tumor gene expression data typically has the characteristics of small samples, high-dimensional, and there are a lot of noise in the original data. So it will spend a lot of time and reduce the effectiveness of classification when using a classifier to forecast a new sample [4]. Based on this, exploring a reasonable and efficient feature gene selection method plays a key role in diseases diagnosis and diseases prediction [5].

Data clustering and classification are important methods of data mining, and they are also the most important tools to analyze the gene expression profiles and identify gene function [4–6]. Due to the high-dimensional and small sample problems of the gene-chip, an optimization algorithm is required to select a gene subset having best disease recognition ability from the attributes. That is, the selected gene subset plays an important role in the process of cancer identification [7,8]. Currently feature selection methods can be divided into three categories: filters, wrappers, and embedded methods [4,9,10]. Particularly, both wrappers and embedded methods consider the correlativity between genes, thus the feature genes select-

*Address for correspondence: Jiucheng Xu, College of Computer and Information Engineering, Henan Normal University, Xinxiang, China. Tel.: 0373-3329075; Fax: 0373-3329075; E-mail: xjc@htu.cn.

ed by the two methods are more interpretable. Literatures [11,12] pointed out, a good feature selection algorithm should be reasonable and efficient, and can find the typical genome containing fewer genes. So the selected subset of attributes and decision class not only have a stronger correlativity between attributes, but also have a smaller redundancy. To remove redundancy effectively, Wang et al. [13] proposed a heuristic width priority search algorithm which looks for information using the classification performance of SVM as the evaluation criterion to eliminate redundant genes; Chuang et al. [14] combined the binary particle swarm optimization algorithm with genetic algorithm, and the k neighbor classifier was used to reduce the redundant genes. The above methods have solved the negative effects caused by the redundancy to some extent. However, it is easy to cause over-fitting and poor generalization performance in the gene selection process. Therefore, a machine learning method with strong robustness needs to be put forward. The unsupervised learning is based on a certain evaluation criterion, and it looks for a feature subset that can be better explain the natural classification of the data; while the supervised learning uses feature selection class label to directly partition, and it chooses subset with strong correlation or some kind of low classification error rate as the optimal feature subset. Consequently, we select a feature subset from all characteristics based on the feature selection. The logistic regression model is a linear regression model, which overcomes the insufficiency of traditional methods on the selection model, and avoids the information loss caused by data discretization. It is widely used in the analysis of gene expression profiles. Aiming to avoid the over-fitting in sample data and model, the correlation information entropy is adopted to weed out redundant genes.

The structure of the rest of this paper is as follows: Section 2 introduces the concepts of binomial logistic regression model and correlation information entropy. An effective and efficient feature selection method based on logistic and correlation information entropy is proposed in Section 3. To evaluate the performance of the proposed algorithm, applying it to two gene expression data sets, and comparing it with other three algorithms. The experimental results are presented in Section 4. Finally, the conclusion is drawn in Section 5.

2. Basic concepts

2.1. Binomial logistic regression model

The binomial logistic regression model [15] is a classification model, which can be represented by the conditional probability distribution of $P(Y|X)$ in the form of parameterized logistic distribution, and the binomial logistic regression model can be defined as $P(Y=1|x)=\frac{exp(w\cdot x)}{1+exp(w\cdot x)}$, $P(Y=0|x)=\frac{1}{1+exp(w\cdot x)}$. Here, $x\in R^n$ is the input variable, $Y\in\{0,1\}$ is the output variable. Further on, the $w\cdot x$ is inner product of w and x, in which $w=(w^{(1)},w^{(2)},\ldots,w^{(n)},b)^T$, $x=(x^{(1)},x^{(2)},\ldots,x^{(n)},1)^T$, while w is the weight vector, and b is offset. If the event occurs with probability p, the probability of the event can be obtain as $\frac{p}{1-p}$. Hence, the logarithmic probability or logit function of the event is $logit(p)=log\frac{p}{1-p}$. The two conditional probabilities values will be compared in the logistic regression, partition instance x into the category that has a larger probability value.

2.2. Information entropy

Suppose *X* is a discrete random variable, the probability density function is $p(x)$, and then the uncertainty degree of the variable *X* can be represented by information entropy $H(x)$, which is denoted

by $H(x) = -\sum_{x \in X} p(x) log p(x)$. Clearly, the entropy is the distribution function of random variable *X*, which depends on its probability distribution instead of the actual value of *X*, and this can avoid the interference of noise samples to a certain extent. The greater of the value of information entropy is, the higher the uncertainty degree of *X* is. Giving a probability distribution $P = (P_1, P_2, \ldots, P_n)$, the information entropy carried by the distribution is called P entropy, and the formula is defined as $I(p) = -(p(x_1) \times log_2 p(x_1) + p(x_2) \times log_2 p(x_2) + \ldots + p(x_n) \times log_2 p(x_n))$. When the order of variables $x_1, x_2, \ldots, x_n$ changes, the entropy value remains unchanged. In other words, the entropy value only depends on the overall distribution probability of sample values.

3. Feature gene selection method based on logistic and the correlation information entropy

3.1. Binomial logistic regression model

The logistic regression model only deals with numeric variables, and classification variables are assigned 1 or -1, the value 1 denotes positive samples, the value -1 denotes negative samples, logistic regression compares two conditional probability values, divides instance X into the class having larger probability value [15]. Based on statistical software SPSS, logistic regression on dataset gene respectively calculated square values and P values of all gene variables, because the first screening of variables only deletes small chi-square value, input and output of condition variables should not be too strict. Here, set the threshold $P = 0.3$. If the P is more than 0.3, exclud this variable. As we know, the gene having high estimation has a higher ability of information classification, because the gene is associated with other genes in the data set. Based on the estimation, logistic regression is used to select genes in this paper.

3.2. Correlation information entropy

Correlation information entropy can measure the correlation between multiple variables. Suppose that the multiple variable and nonlinear system *S* has *Q* variables, and the multivariate time series matrix of this system is P at moment $t(t = 1, 2, \ldots, K)$, where $P \in \mathbf{R}^{K \times Q}$, and $y_i(t)$ denotes the value of t at the time i. In general, $Q \ll K$, having $P = y_i(t)$, where $1 \leq t \leq K, 1 \leq i \leq Q$, there exists correlation coefficient matrix R, where $R \in \mathbf{R}^{Q \times Q}$, $R = P^T \cdot P$. Hence, it can be transformed into $R = \begin{bmatrix} 1 & r_{12} & \cdots & r_{1Q} \\ r_{21} & 1 & \cdots & r_{2Q} \\ \vdots & \vdots & \ddots & \vdots \\ r_{Q1} & r_{Q2} & \cdots & 1 \end{bmatrix}$.

Definition 1. Suppose gene number is N, the gene number of feature gene subset is W. There is eigenvalue λ_j in the correlation coefficient matrix, where $\lambda_j > 0$, $j = 1, 2, \cdots, W$, and $W \ll N$. The correlation information entropy of feature gene is defined as $H_R = -\sum_{j=1}^{W} \frac{\lambda_j}{W} log_W \frac{\lambda_j}{W}$.

The greater the H_R is, the bigger the correlation information entropy is. The correlation of genes selected from attribute set is smaller, namely, the independent is bigger.

3.3. Logistic correlation information entropy algorithm

To effectively remove the redundant genes, the correlation information entropy is put forward in this paper, which can search feature genes accurately in smaller feature gene sets. We know that the corre-

lation coefficient matrix of a random variable reflects degree of correlation among variables. Through analyzing the linear correlation of n random variables $x_1, x_2, \cdots x_n$, it can be measured by mean square error E: $E = \alpha^T R\alpha = x^T \wedge x = \lambda_1 x_1^2 + \lambda_2 x_2^2 + \cdots + \lambda_n x_n^2 \geq 0$. When a linear combination of variables is an ordinary coefficient equation, the size of the E is decided by the eigenvalue $\lambda_1, \lambda_2, \cdots, \lambda_n$, that is, the eigenvalues of correlation matrix reflect the variables' degree of the linear correlation.

Definition 2. Given the information entropy of feature gene named H_R, then, the selected feature gene subset F with maximum information entropy is defined as $Max\ H_R(s \cup g_i), i = 1, 2, 3, \cdots, n$. Where, g_i is the gene variable; and n is the number of genes.

Combine the above analysis, the specific algorithm proposed in this paper depicted as follows:

Algorithm 1 Logistic correlation information entropy algorithm $(Lciea)$

Input: The train data set TR, test data set TE, Relief filter values δ, $S = \{g_1, g_2, \cdots, g_n\}$;
Output: Feature gene set F;

```
F = null; H_R = null; // the initial state is empty;
Relief (TR) // using relief algorithm for feature assignments;
Get feature weight w = {w_1, w_2, ..., w_n};
for i = 1, 2, ..., n do
    if (w_i > δ) F = F ∪ {g_i}; // put g_i into F and get the new feature set F;
    Sorting feature gene weights of F from the largest to the smallest to get F_s = g_1, g_2, ..., g_m;
end for
F= null; // initialize empty set;
for i = 1, 2, ..., m do
    Calculate H_R(F ∪ {g_i}); // add feature gene and calculate its correlation information entropy;
    if (H_R(F ∪ {g_i}) − H_R(F)) > 0 then
        F = F ∪ {g_i}; // if the correlation information entropy increases, add the gene to F;
    end if
    if (H_R(F ∪ {g_i}) − H_R(F)) < 0 then
        F = F − g_i; // otherwise, remove the gene;
    end if
    Update the F;
end for
return Feature gene set F.
```

4. Experiment analysis

4.1. The experimental data description

In order to verify the validity of the algorithm, this paper adopts breast cancer database set and gastric cancer data set as the experimental data which from the UCI. In the experiments, two canner recognition data sets are collected to test the performances of Lciea. The breast cancer data set consists of 84 samples and 9216 gene expression data and the gastric cancer has 40 samples and 1520 gene expression data. For gastric cancer data set with 1520 genes, logistic stepwise regression of the 40 goals is carried out, and the number of the variables are preliminary reduced to 942; the breast cancer data set with 9216

genes, logistic stepwise regression of the 84 goals is carried out, and the number of the variables are preliminary reduced to 5623. After the above algorithm, two feature subsets denoted by S_{breast} and $S_{gastric}$ are obtained. In allusion to the cancer gene expression data sets, LIBSVM is used as classifier. And using RBF as kernel function, the penalty factor C =100. Other parameters are the current values by default. In the process of removing irrelevant genes, Relief is only used to descending order according to the weight of all genes in the first stage of feature selection algorithm based on the traditional Relief, which meets the goals of removing the irrelevant genes. In this paper, genes having greater influence on the classification are obtained which is based on logistic regression model, then using Relief algorithm to remove irrelevant genes. Through the experiment, the gene classification weight and the scatter of genetic classification weight can be obtained respectively as shown in Figures 1–4 and Figures 5–8.

Figures 1 and 5 and Figures 3 and 7 are the gene classification weight column graphs obtained by the traditional Relief algorithm, while Figures 2 and 6 and Figures 4 and 8 are the gene classification weight column graphs obtained by the Lciea algorithm. From the figures, when the weights of classification gene are equal, the gene number of Figures 2 and 6 is smaller than gene number in Figures 1 and 5, and the total number of genes having larger classification weight in Figures 2 and 6 is less than the total number of genes having larger classification weight in Figures 1 and 5; the gene number of Figures 4 and 8 is smaller than gene number in Figures 3 and 7, and the total number of gene having larger classification weight in Figures 4 and 8 is less than the total number of genes having larger classification weight in Figures 3 and 7. For example, when weight is less than 500, in breast data set, there are about 4000 genes

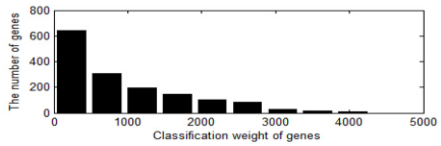

Fig. 1. Gastric gene classification weight column by Relief.

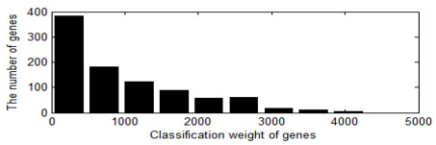

Fig. 2. Gastric gene classification weight column by Lciea.

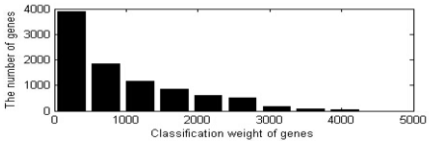

Fig. 3. Breast gene classification weight column by Relief.

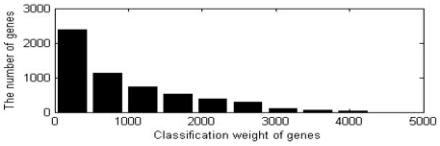

Fig. 4. Breast gene classification weight column by Lciea.

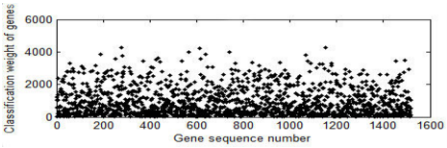

Fig. 5. Gastric gene classification weight scatterplot by Relief.

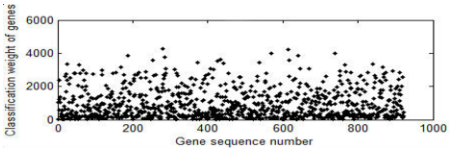

Fig. 6. Gastric gene classification weigh scatterplot by Lciea.

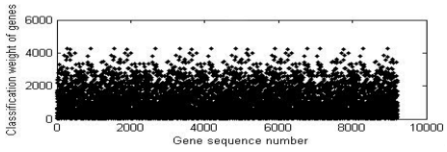

Fig. 7. Breast gene classification weight scatterplot by Relief.

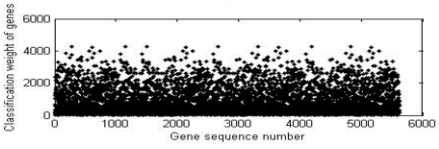

Fig. 8. Breast gene classification weigh scatterplot by Lciea.

before logistic regression, while there are about 2500 genes after using this algorithm. For this reason, the logistic regression significantly reduce the number of genes, the genes that have a greater impact on classification are obtained.

4.2. Experimental result analysis

In this paper, three redundant genes eliminating methods are studied, including Recorre feature selection algorithm by using mutual information method to remove redundant genes, Resbsw feature selection algorithm by using backward search method to remove redundant genes and Ners (Neighborhood rough set) feature selection algorithm by using attribute reduction to remove redundant genes. The proposed algorithm in this paper uses correlation information entropy to remove redundant genes. In order to compare classification performance of four kinds of the algorithms, the 10-fold cross-validation is used to evaluate the method. Then, the average of classification result is calculated as final classification results of each algorithm. In Tables 1 and 2, B stands for breast data set; G stands for gastric data set.

A three-step approach is presented in this paper: firstly, applying logistic regression to select a preliminary set; secondly, using Relief to rank the genes; finally, building the subset with information entropy. In Table 1, the effectiveness of each step in building the final feature subset is analyzed. For example, in the step of logistic regression, 942 features with classification accuracy of 86.2431% from breast data set and 5623 features with classification accuracy of 88.1941% from gastric data set are obtained; in the step of Relief, 639 features with classification accuracy of 88.3687% from breast data set and 214 features with classification accuracy of 89.8165% from gastric data set are achieved. As we can see, the Lciea algorithm consistently outperforms the other algorithms. It can reach its lowest value namely 410 genes with the highest classification accuracy of 96.3928% were selected from breast data set and 76 genes with the highest classification accuracy of 95.6243% were selected from gastric data set.

From Table 2, the results show the Lciea algorithm is superior to Resbsw method and Recorre method in the two aspects of the feature gene number and the classification accuracy. Although Recorre method can obtain the smaller feature subset, its classification accuracy is lower and time complexity is larger, and although the classification accuracy of Resbsw method is higher than the Recorre method, its feature subset is larger. In breast data set, 683 feature genes obtained by the Recorre method is less than 754 features obtained by Resbsw method, but the former classification accuracy is 84.6243%, which is lower than the latter classification accuracy of 86.5814%. While the classification accuracy of Ners method is worst and the time complexity is $O(m \times nlogn)$, m is the number of samples and n is the number of

Table 1

Classification performance comparison of each step in building the Lciea algorithm.

Algorithms	Feature subset		Accuracy	
	B	G	B	G
logistic regression	942	5623	86.2431%	88.1941%
Relief	639	214	88.3687%	89.8165%
Lciea	410	76	96.3928%	95.6243%

Table 2

Classification performance comparison of the three algorithms.

Algorithms	Feature subset		Accuracy		Time complexity	
	B	G	B	G	B	G
Ners	421	95	80.4325%	83.5179%	$O(m \times nlogn)$	$O(m \times nlogn)$
Recorre	683	112	84.6243%	85.6217%	$O(n^2)$	$O(n^2)$
Resbsw	754	134	86.5814%	86.9712%	$O(n)$	$O(n)$
Lciea	410	76	96.3928%	95.6243%	$O(n)$	$O(n)$

features. By the analysis above, the Lciea algorithm not only can obtain the least genes in feature gene set, but can get a big promotion in the classified accuracy. This algorithm can select 410 feature genes with classification accuracy of 96.3928% and the time complexity is $O(n)$, so this method can eliminate redundant genes effectively.

5. Conclusion

In this paper, the logistic regression model and information entropy is introduced into the feature gene selection algorithm. The method has two advantages: first, the Lciea algorithm can obtain fewer feature genes; second, the Lciea algorithm can achieve higher classification accuracy without any increase of time complexity.

Acknowledgements

The work is supported by National Natural Science Foundation of China (61370169, 61402153), Project of Henan Science and Technology Department (142102210056), Project of Henan Educational Department (12A520027, 13A520529), and Youth Key Teachers of Henan Normal Univ.

References

[1] C.D. Qin, S. Liu and S.F. Zhang, Method for extracting the tumor gene based on the support vector machine, Journal of Xidian University **39** (2012), 192–196.

[2] F. Liu, Time-lagged Co-expression Gene Analysis Based on Biclustering Technology, Biotechnology & Biotechnological Equipment **27** (2013), 4031–4039.

[3] F. Liu and L.B. Wang, Biclustering of time-lagged gene expression data using real data, Journal of Biomedical Science and Engineering **3** (2010), 217–220.

[4] J.C. Xu, L. Sun and Y.P. Gao, An ensemble feature selection technique for cancer recognition, Bio-Medical Materials and Engineering **24** (2014), 1001–1008.

[5] B. Scholkopf, K. Tsuda and J-P. Vert, Kernel Methods in Computation Biology, MIT Press, Cambridge, 2004, pp. 299–318.

[6] J. Li, X.L. Tang and Y.D. Wang, Research on Gene Expression Data Based on Clustering classification Technology, Chinese Journal of Biotechnology **21** (2005), 668–673.

[7] E. Lotfi and A. Keshavarz, Gene expression microarray classification using PCA-BEL, Computers in Biology and Medicine **54** (2014), 180–187.

[8] H. Banka and S. Dara, A Hamming distance based binary particle swarm optimization (HDBPSO) algorithm for high dimensional feature selection, classification and validation, Pattern Recognition Letters **52** (2015), 94–100.

[9] R. Kohavi and G. John, Wrapper for feature subset selection, Artificial Intell **97** (1997), 273–324.

[10] L. Sun, J.C. Xu and Y. Tian, Feature selection using rough entropy-based uncertainty measures in incomplete decision systems, Knowledge-Based Systems **36** (2012), 206–216.

[11] F. Zhou and J.Y. He, Survey of the Gene Selection Technologies Based on Microarray in Bioinformatics, Computer Science **34** (2007), 143–150.

[12] J. Guan, F. Han and S.X. Yang, Gene Selection Algorithm Based on Particle Swarm Optimization and J-divergence Entropy Information, Computer Engineering **39** (2013), 188–190.

[13] S.L. Wang, J. Wang and H.W. Chen, Heuristic Breadth-First Search Algorithm for Informative Gene Selection Based on Gene Expression Profiles, Chinese Journal of Computers **31** (2008), 636–649.

[14] L.Y. Chuang, C.H. Yang and J.C. Li, A Hybrid BPSO-CGA Approach for Gene Selection and Classification of Microarray Data, Journal of Computational Biology **19** (2012), 68–82.

[15] H. Li, Logistic regression model, in: Statistical learning method, H. Xue, ed., Tsinghua University Press, Beijing, 2012, pp. 77–80.