African Journal of Biotechnology Vol. 11(45), pp. 10438-10445, 5 June, 2012
Available online at http://www.academicjournals.org/AJB
DOI:10.5897/AJB11.1016
ISSN 1684–5315 © 2012 Academic Journals

Full Length Research Paper

An entropy-based improved k-top scoring pairs (TSP) method for classifying human cancers

Chunbao Zhou[1,2], Shuqin Wang[3], Enrico Blanzieri[4] and Yanchun Liang[1]*

[1]College of Computer Science and Technology, Jilin University, Key Laboratory of Symbol Computation and Knowledge Engineering of the Ministry of Education, Changdhun 130012, People's Republic of China.
[2]Supercomputing Center, Computer Network Information Center, Chinese Academy of Sciences, Beijing, 100190, China.
[2]College of Computer and Information Engineering, Tianjin Normal University, Tianjin 300387. China.
[3]Department of Information Engineering and Computer Science, University of Trento, Via Sommarive 14, 38050 – Povo (TN) Italy.

Accepted 14 February, 2012

Classification and prediction of different cancers based on gene-expression profiles are important for cancer diagnosis, cancer treatment and medication discovery. However, most data in the gene expression profile are not able to make a contribution to cancer classification and prediction. Hence, it is important to find the key genes that are relevant. An entropy-based improved k-top scoring pairs (TSP) (Ik-TSP) method was presented in this study for the classification and prediction of human cancers based on gene-expression data. We compared Ik-TSP classifiers with 5 different machine learning methods and the k-TSP method based on 3 different feature selection methods on 9 binary class gene expression datasets and 10 multi-class gene expression datasets involving human cancers. Experimental results showed that the Ik-TSP method had higher accuracy. The experimental results also showed that the proposed method can effectively find genes that are important for distinguishing different cancer and cancer subtype.

Key words: Cancer classification, gene expression, k-TSP, information entropy, gene selection.

INTRODUCTION

Classification and prediction of different cancers based on gene-expression profiles have recently received a great deal of attention in the field of bioinformatics (Golub et al., 1999; Khan et al., 2001; Hedenfalk et al., 2001). Recently, many gene selection methods have been developed: a method of key genes selection using SVM by Guyon et al. (2002); a method of classifying cancers to specific four distinct diagnostic categories based on their gene expression signatures using artificial neural networks (ANNs), a sensitivity analysis method to find key genes by Khan et al. (2001); and a method of finding key genes using Bayesian method by Zhou et al. (2004).

At present, there are many machine learning methods for the classification of cancers. Li and Ruan, (2005) first used the exponential of classification information of gene for data pre-processing by selecting the genes with classification characters and removing the unconcerned ones. Then they used support vector machine (SVM) to predict cancer classes. Li et al. (2002) also introduced an unsupervised gene filtering algorithm to reduce the data noise of subtype calculation. First, they presented a probabilistic model for classification in the sample, and then they used the relative entropy method to acquire the genes with the greatest classification contribution as key genes based on the results of the gene cluster. Finally, they applied the cluster of key genes to the classification of cancers. Liu et al. (2006) predicted the category of the testing sample data by comparing an expansive topology map with the primitive topology map, and finding the two most similar topology maps by using the characteristic information of the topological map. In addition, Zhou et al. (2004) presented a logarithm recursive method to classify cancers, while Helman et al. (2004) used a Bayesian network method. Furthermore, Geman et al. (2004) presented a classification approach completely based on

*Corresponding author. E-mail: ycliang@jlu.edu.cn. Fax: +86 431 85168752.

the top scoring pair (TSP) with relative value of the gene expression. They first calculated the probability for every two genes in various category samples that the gene expression level value of one gene is higher than the other; then they chose one pair of genes with the highest probability. The TSP classifier is an entirely data-driven machine learning approach without any parameter. The classification rules of TSP contain only a pair of genes. Moreover, because in some datasets the classification rules of TSP classifier will change with the addition or deletion of training samples, Tan et al. (2005) proposed an improved method named "k-TSP method" for binary class datasets and a HC-k-TSP scheme for dealing with multi-class datasets. These methods choose the k disjoint top scoring pairs of genes as decision rules rather than only the highest pair and both methods need to calculate the score of each gene pair. However, cancer datasets have a huge size (the datasets considered in this paper contained at least 2,000 genes), thus the two algorithms suffer both high time and space consumption. Also, the presence of genes which are irrelevant to cancer classification influences the accuracy of classification. These limitations show the importance of key genes selection for k-TSP and HC-k-TSP method.

In order to evaluate the ability of Ik-TSP method in cancer classification, we took 9 binary class gene expression datasets and 10 multi-class gene expression datasets, which were used by Tan (2005), as our experimental datasets. We then compared our method with five other machine learning methods. It was demonstrated that the proposed Ik-TSP method obtained an average of 96.28% accuracy, which was the best classifier in 9 binary class datasets and 87.32% accuracy which was the third best one in 10 multi-class datasets. We also compared our method with k-TSP method based on other feature selection methods, and it was found that Ik-TSP method obtained the best accuracy.

METHODS

The k-TSP method

Assuming that a gene expression profile consists of expression values of P genes $\{1,2,...,P\}$ and there are N profiles $X_1, X_2,..., X_N$ available for training. $(y_1, y_2, \cdots, y_N)$ is the vector of class labels for the N samples, where $y_n \in c = \{C_1, C_2, \cdots, C_m\}$, which is the set of possible class label. For example, C_1 refer to the normal tissues and C_2 to the cancer tissues. Tan et al. (2005) replaced the expression values $X_{i,n}$ by their ranks $R_{i,n}$ for comparing their relationship as shown in the following formulas. The aim was to find the 'marker gene pairs' (i, j) $i, j \in \{1,2,...,P\}$ and $i \neq j$, which have a significant difference in the probability of the event $\{R_i < R_j\}$ across the N samples from class C_1 and C_2. Here, the quantities of interest are:

$$p_{ij}(C_m) = \Pr ob(R_i < R_j \mid Y = C_m), m \in \{1,2\} \quad (1)$$

Using Δ_{ij} to represent the 'score' of the gene pair (i, j), Δij can be defined as:

$$\Delta_{ij} = \mid p_{ij}(C_1) - p_{ij}(C_2) \mid \quad (2)$$

Hence, we can compute the 'average rank difference' r_{ij} in class C_m, which is defined as:

$$r_{ij}(C_m) = \frac{\sum_{n \in C_m} (R_{i,n} - R_{j,n})}{\mid C_m \mid}, \quad m \in \{1,2\} \quad (3)$$

Where, $|C_m|$ denotes the number of samples in the class C_m. The 'rank score' of the gene pair (i, j) is then defined as:

$$\Gamma_{ij} = |r_{ij}(C_1) - r_{ij}(C_2)| \quad (4)$$

Tan et al. (2005) selected the gene pairs with the largest rank score among the gene pairs with the score $\Delta_{\max}$. Given a new profile x_{new}, the y_{new} is predicted according to the t-th single gene pair (i, j) as follows. If $p_{ij}(C_1) > p_{ij}(C_2)$ then:

$$y_{new} = h_t(x_{new}) = \begin{cases} C_1, & \text{if } R_{i,\text{new}} < R_{j,\text{new}} \\ C_2, & \text{otherwise} \end{cases} \quad (5)$$

Otherwise, the decision rule is reversed. Tan et al. (2005) also employed an unweighted majority voting procedure based on k top scoring disjoint pairs of genes to compute the class of the new sample. The predictive formula is defined as follows:

$$y_{new} = h_k - TSP(x_{new}) = \arg\max_{c = C_1, C_2} \sum_{t=1}^{k} I(h_t(x_{new}) = c) \quad (6)$$

Where:

$$I(h_t(x_{new}) = c) = \begin{cases} 1 & \text{if } h_t(x_{new}) = c \\ 0 & \text{otherwise} \end{cases}, \quad c \in \{C_1, C_2\} \quad (7)$$

Since the k-TSP method is essentially an exhaustive algorithm for gene pairs, the time and space consumption quadratically increase in the number of genes, even using pruning algorithm (Tan et al., 2005). Frequently, the final number of genes that are actually used in the classification is very small in most classification methods compared with the large number of genes in the datasets. This observation suggests that there are a large number of genes which are not relevant to cancer classification in the dataset; hence gene selection is very important as this could increase the accuracy and decrease the time and space requirement of the classification. Therefore, this study proposed an improved k-TSP method (referred to as Ik-TSP method) that uses the information entropy to select the key genes in the gene expression data. Some classification methods, example SVMs, TSP and k-TSP, are designed only for binary class datasets. Tan et al. (2005) investigated the performance of the k-TSP classifiers for three

different classification schemes for m classes. These are the One-vs.-Others (1-vs-r) scheme, the One-vs.-One (1-vs-1) scheme and the hierarchical classification (HC) scheme. The hierarchical classification (HC) scheme was used for k-TSP method in this study.

The improved k-TSP method based on Information entropy

Information entropy has been widely used in the field of classification (Shannon et al., 2003; Wei et al., 2004). Since there are a lot of uncertain factors involved in the information contained in the data, we selected the attributes that contained large amount of certain information in the classification, as they are specially important for classification. Information Entropy reflects the content of certain information contained in an attribute. The definition of the information entropy is as follow:

$$H = -\sum_i p_i \log(p_i) \quad (8)$$

Where, p_i denotes the probability of the object in the i-th class in all the samples, which generally can be estimated by the frequency of the object in that class in the dataset. The greater certain information is, the smaller the information entropy.

For cancer classification, the information provided by each gene for the classification is different. In fact, some genes do not provide any information for classification, while others could determine the class of the sample; hence the latter genes are important and called key genes. Our method used the information entropy to select a group of key genes. The method removes the genes which contain little information, by choosing a set of genes whose information entropy is small. This method not only reduces the amount of data for classification, but also avoids the negative effects on the classification performance due to equal treatment of genes with different information content. Accordingly, the accuracy will be improved.

For a discrete attribute, it is easy to calculate the information entropy while the variable takes different values for classification. However, for a continuous attribute, it is hard to calculate the information entropy for classification, while the variable takes every possible value. The expression value of each gene in the dataset of gene expression profiles for cancer classification is continuous, so each attribute need to be discretized. In this paper, the discretization is included in the algorithm for the key genes selection method based on information entropy described as follows:

Input: S training samples, N genes and threshold H.
Output: A set θ of key genes.
1. For i =1 to N step 1 do

a. Make an order list O of samples according to the expression value of the i-th gene in all samples from largest to smallest, and set maximal entropy E_i.
b. Repeat m times: (m = S - 2n + 1, n is the number of samples in minimal set of samples, n = 3 in this study)

(1)Select first n samples from list O, calculate entropy h_1 for gene i against the n samples.
(2)Calculate entropy h_2 for gene i against all samples except for the n samples.
(3) $H_i = (n/S)h_1 + ((S-n)/S)h_2$.
(4) If $H_i < E_i$, then $E_i = H_i$
(5) n=n+1.

2. Set an entropy threshold H.
3. Make a list θ, if $E_i < H$ for gene i, then add gene i into θ.
4. Return list θ.

First, the expression values are arranged in descending order for each gene i in the gene expression profile. Secondly, the first three samples are selected as a group to calculate the corresponding information entropy, and then the information entropy is calculated using the remaining samples and sum up them. Next, the first four samples are chosen as a group to calculate the corresponding information entropy, and the information entropy is calculated using the remaining samples and sum up them, for S-2n+1 times. Finally, a threshold for the entropy is set, and a set of genes is constructed, which consists of all the genes whose information entropy is less than the threshold to be used as key genes set by k-TSP and HC-k-TSP method.

We selected different key genes sets based on the different entropy threshold ranges for different datasets. Since there are both a testing set and a training set for each multi-class gene expression dataset, but there is only a training set for each binary class gene expression dataset, we used different ways of selection of the key genes for multi-class and binary class gene expression datasets. For binary class gene expression datasets, we selected key genes depending on the dataset obtained leaving one sample out from the whole training dataset. Whereas for multi-class gene expression datasets, key genes are selected depending on the whole training dataset.

This meant that we used key genes selection method based on information entropy only once for multi-class dataset and many times for binary class gene expression dataset.

Microarray data

In order to validate the effectiveness of the Ik-TSP method, we compared Ik-TSP method with other machine learning methods and k-TSP method based on other feature selection methods. We used data downloaded from Tan et al. (2005). There are 9 binary class gene expression datasets and 10 multi-class gene expression datasets involving human cancers. The information about the datasets is shown in Tables 1 and 2.

Other machine learning methods and feature selection methods

We compared the performance of the Ik-TSP classifier with some well-known machine learning methods for cancer classification including C4.5 decision trees (DT), Naive Bayes (NB), k-nearest neighbor (k-NN), support vector machines (SVM) and prediction analysis of microarrays (PAM). The results for these five methods are available from Tan et al. (2005).

The Ik-TSP method is an improved k-TSP method based on information entropy feature selection method. In order to compare our feature selection method with other feature selection methods, we compared the Ik-TSP method with the k-TSP method based on other feature selection methods, including Relief, Sequential Floating Forward selection method (sffs) and Sequential Forward Selection method (sfs) (Acu, 2003). These compared methods are the functions in dprep package based on R language (http://cran.stat .ucla.edu/src/contrib/Archive/dprep/).

RESULTS

Comparison with machine learning methods

Here, the Leave-One-Out Cross-Validation (LOOCV) was

Table 1. Binary class gene expression datasets.

Dataset	No. of genes (P)	No. of samples (N)		Reference
		C_1	C_2	
Colon	2000	40 (T)	22 (N)	(Alon et al., 1998)
Leukemia	7129	25 (AML)	47 (ALL)	(Golub et al., 1998)
CNS	7129	25 (C)	9 (D)	(Pomeroy et al., 1998)
DLBCL	7129	58 (D)	19 (F)	(Shipp et al., 1998)
Lung	12533	150 (A)	31 (N)	(Gordon et al., 1998)
Prostate1	12600	52 (T)	50 (N)	(Singh et al., 1998)
Prostate2	12625	38 (T)	50 (N)	(Stuart et al., 1998)
Prostate3	12626	24 (T)	9 (N)	(Welsh et al., 1998)
GCM	16063	190 (C)	90 (N)	(Ramaswamy et al., 1998)

Table 2. Multi-class gene expression datasets.

Dataset	Number of classe	Number of gene (P)	Number of sample (N)		Reference
			Training	Testing	
Leukemia1	3	7129	38	34	(Golub et al., 1999)
Lung1	3	7129	64	32	(Beer et al., 2002)
Leukemia2	3	12582	57	15	(Armstrong et al., 2002)
SRBCT	4	2308	63	20	(Khan et al., 2001)
Breast	5	9216	54	30	(Perou et al., 2000)
Lung2	5	12600	126	67	(Bhattacharjee et al., 2001)
DLBCL	6	4026	58	30	(Alizadeh et al., 2000)
Leukemia3	7	12558	215	112	(Yeoh et al., 2002)
Cancers	11	12553	100	74	(Su et al., 2001)
GCM	14	16063	144	46	(Ramaswamy et al., 2001)

Table 3. LOOCV accuracy of classifiers for binary class gene expression datasets (%).

Method	Leukemia	CNS	DLBCL	Colon	Prostate1	Prostate2	Prostate3	Lung	GCM	Average
TSP	93.80	77.90	98.10	91.10	95.10	67.60	97.00	98.30	75.74	88.26
k-TSP	95.83	**97.10**	97.40	90.30	91.18	75.00	97.00	98.90	85.40	92.36
I*k*-TSP	**100**	94.12	**98.70**	**93.55**	**98.04**	**90.91**	**100**	**99.45**	91.79	**96.28**
DT	73.61	67.65	80.52	80.65	87.25	64.77	84.85	96.13	77.86	79.25
NB	**100**	82.35	80.52	80.65	87.25	73.85	90.91	97.79	84.29	81.17
k-NN	84.72	76.47	84.42	74.19	76.47	69.32	87.88	98.34	82.86	81.63
SVM	98.61	82.35	97.40	82.26	91.18	76.14	**100**	**99.45**	**93.21**	91.18
PAM	97.22	82.35	85.71	85.48	91.18	79.55	**100**	**99.45**	79.29	88.91

The best prediction rate for each dataset was highlighted in boldface.

employed to estimate the classification error rate for binary class dataset. For multi-class dataset, the accuracy for the testing dataset was directly the result. In Table 3, the Ik-TSP method was the best as it performed an average of 96.28% in the LOOCV accuracy, followed by k-TSP (92.36%), SVM (91.18%) etc. Ik-TSP outperformed k-TSP in eight cases (Leukemia, DLBCL, Colon, Prostate1, Prostate2, Prostate3, Lung and GCM), while k-TSP outperformed Ik-TSP in one cases (CNS). In Table 4, the proposed Ik-TSP method showed the third best performance over the 10 datasets, but did not exceed the 1-vs-1-SVM (88.11%) and PAM (88.50). The Ik-TSP achieved an average accuracy of 87.32%. Compared with HC-k-TSP, the Ik-TSP was superior in five cases (Leukemia1, Lung1, Breast, DLBCL and Cancer), inferior in three cases (Lung2, Leukemia3 and GCM) and the same in two cases (Leukemia2 and SRBCT).

Table 4. Accuracy of classifiers for multi-class gene expression datasets (%).

Method	Leu1	Lung1	Leu2	SRBCT	Breast	Lung2	DLBCL	Leu3	Cancers	GCM	Average
TSP	97.06	71.88	80.00	95.00	66.67	83.58	83.33	77.68	74.32	52.17	78.17
k-TSP	97.06	78.13	**100**	**100**	66.67	94.03	83.33	82.14	82.43	**67.39**	85.12
I*k*-TSP	**100**	84.38	**100**	**100**	83.33	92.54	93.33	65.18	**89.19**	65.22	87.32
DT	85.29	78.13	80.00	75.00	73.33	88.06	86.67	75.89	68.92	52.17	76.35
NB	85.29	81.25	**100**	60.00	66.67	88.06	86.67	32.14	79.73	52.17	73.20
k-NN	67.65	75.00	86.67	30.00	63.33	88.06	93.33	75.89	64.86	34.78	67.96
SVM	79.41	**87.50**	**100**	**100**	83.33	97.01	**100**	84.82	83.78	65.22	88.11
PAM	97.06	78.13	93.33	95.00	**93.33**	**100**	90.00	**93.75**	87.84	56.52	**88.50**

The best prediction rate for each dataset was highlighted in boldface.

Table 5. LOOCV accuracy of feature selection methods for binary class gene expression datasets (%).

Method	Leukemia	CNS	DLBCL	Colon	Prostate1	Prostate2	Prostate3	Lung	GCM	Average
I*k*-TSP	100	94.12	98.70	93.55	98.04	90.91	100	99.45	91.79	96.28
Relief	97.22	97.06	98.70	91.94	94.12	89.77	96.97	99.45	92.14	95.19
sffs	18.31	0	26.32	31.15	53.47	0	0	16.11	65.97	23.48
sfs	28.17	100	72.37	42.62	50.01	32.20	0	81.11	75.32	53.53

Table 6. Accuracy of feature selection methods for multi-class gene expression datasets (%).

Method	Leu1	Lung1	Leu2	SRBCT	Breast	Lung2	DLBCL	Leu3	Cancers	GCM	Average
I*k*-TSP	100	84.38	100	100	83.33	92.54	93.33	65.18	89.19	65.22	87.32
Relief	82.35	65.63	93.33	95.00	73.33	77.61	90.00	84.82	83.78	60.87	80.67
sffs	91.18	65.63	80.00	55.00	56.67	79.10	56.67	50.00	35.14	26.09	59.55
sfs	79.41	56.25	26.67	50.00	56.67	85.07	60.00	41.07	20.27	23.91	49.93

Comparison with feature selection methods

In this article, the parameters for different feature selection methods were described as follows. For relief, the cut-off point to select the features was from 0.01 to 0.1. For sffs, the classifier was k nearest-neighbors method (knn), in which the number of nearest neighbors was 5 and the number of repetitions was 5. For sfs, the classifier was also k nearest-neighbors method (knn), in which the number of neighbors to use for the knn classification was 3 and the number of times to repeat the selection was 10. Moreover, from Table 5, we could observe that Ik-TSP (96.28%) was a little better than relief (95.19%) and much better than sffs (23.48%) and sfs (53.53%). There were two reasons for the appearance of 0 in Table 5. One was that the sffs and sfs methods only select one gene from datasets, and the k-TSP method which was based on the gene pairs cannot obtain results; the other was that it is the real result from the k-TSP method. Also, from Table 6 we can see that Ik-TSP (87.32%) was a little better than relief (80.67%) and much better than sffs (59.55%) and sfs (49.93%).

From the above description, it is clear that the feature selection method based on information entropy was best for k-TSP method as it enables the k-TSP method to obtain better accuracy. The accuracy of relief method was a little less than our feature selection method, so it was also a good choice for k-TSP method. The basic idea of relief method was to choose features known as the relevant features that can be most distinguished between classes. At each step of an iterative process, an instance x was chosen at random from the dataset and the weight for each feature was updated according to the distance of x to its near miss and near hit (Kira and Rendel, 1992; Kononenko et al., 1997).

The idea of relief is different from the information entropy. The relief method concerns the genes that can distinguish different classes, while the information entropy method concerns the genes that contain classfication information. The relief method may drop some useful gene pairs for k-TSP method. A concrete analysis will be performed in our future work.

DISCUSSION

Number of genes used in classifier

From the experimental results, it can be observed that

Table 7. Number of genes used in the classifiers for binary class gene expression datasets.

Method	Leukemia	CNS	DLBCL	Colon	Prostate1	Prostate2	Prostate3	Lung	GCM
TSP	2	2	2	2	2	2	2	2	2
k-TSP	18	2	2	18	10	2	2	14	18
I*k*-TSP	22	55	40	31	5	37	16	9	23
DT	2	2	3	3	4	4	1	3	14
PAM	2296	4	17	15	47	13	701	9	47

Table 8. Number of genes used in the classifiers for multi-class expression datasets.

Method	Leukemia1	Lung1	Leukemia2	SRBCT	Breast	Lung2	DLBCL	Leukemia3	Cancers	GCM
TSP	4	4	4	6	8	8	10	12	20	26
k-TSP	36	20	24	30	24	28	46	64	128	134
I*k*-TSP	20	36	28	18	56	72	90	56	180	98
DT	2	4	2	3	4	5	5	16	10	18
PAM	44	13	62	285	4822	614	3949	3949	3338	2008

the Ik-TSP method was superior compared with the classic methods such as PAM and SVM. The number of genes involved in the Ik-TSP method was significantly less than that in PAM in most cases, and obviously less than that in SVM which uses all the genes. NB, k-NN and SVM methods use all the genes for classification, so there are only TSP, k-TSP, Ik-TSP, DT and PAM method in Tables 7 and 8. The maximum value of k for binary classifier in Ik-TSP method is manually selected as 10.

However, small perturbations in the training samples for DT method can lead to large differences in its tree-structure (Dietterich, 2000; Tan and Gilbert, 2003). Hence, the Ik-TSP method is better than the DT, TSP and k-TSP method, although there are many ways to reduce the number of genes through using gene selection methods before training a classifier (Li et al., 2004; Bø and Jonassen, 2002; Dudoit and Fridlyand, 2003).

Biological significance of the Ik-TSP classifier

We only illustrated the rules derived from the Ik-TSP classifier applied to the Leukemia dataset, because Leukemia has been investigated for a long time and there are many pathway data and other data for comparison. By using the Ik-TSP method, we obtained some very important genes, and found their gene names from NCBI (http://www.ncbi.nlm.nih.gov/) for analyzing their important biological signification. For binary class problem, it is Leukemia dataset, and for multi-class problem, it is Leukemia2 dataset.

Many popular approaches such as SVM and some others, have predominantly focused on classification based on all the genes in a dataset, and do not care about the interrelations among genes. One way to address this problem is to look at gene sets rather than all the genes or only one gene. However, how to find the gene set is yet another problem. A number of methods and programs have been developed to solve gene groupings based on Gene Ontology (GO) (Gene Ontology Consortium, 2004). When considering the gene set, we can easily associate it to the pathway because pathways are sets of genes that serve a particular cellular or physiologic function. They are very important for every activity in life, such as biosynthesis, metabolism and so on. Hence, pathway-based methods present another promising approach. It is obvious that focusing pathways relevant to a particular phenotype, e.g. cancer, can help researchers to focus on a few sets of genes. They are particularly useful for generating further biological hypotheses of interest. Although the Ik-TSP method use a gene set for classification, it does not include the interrelations among genes. The interrelations among genes might make the Ik-TSP method getting better accuracy. In this paper, we also analysed the pathway information for genes.

For binary class gene expression dataset

There are 22 genes to distinguish ALL from AML. Among these 22 genes, 3 genes (CD33, Zyxin and CCND3) are in correlation with cancer pathogenesis (Golub et al., 1999). In addition, SPTAN1 is involved in Tight junction. FAT is an ortholog of the Drosophila fat gene, which encodes a tumor suppressor essential for controlling cell proliferation during Drosophila development, and its product is likely to be important in developmental processes and cell communication. APLP2 has been linked with leukemia (Mutis et al., 1999; Yang, 2004). Granulins are a family of secreted, glycosylated peptides that are cleaved from a single precursor protein with 7.5 repeats

of a highly conserved 12-cysteine granulin/ epithelin motif. The 88 kDa precursor protein, progranulin, is also called proepithelin and PC cell-derived growth factor. Cleavage of the signal peptide produces mature granulin which can be further cleaved into a variety of active, 6 kDa peptides. These smaller cleavage products are named granulin A, granulin B, granulin C, etc. Epithelins 1 and 2 are synonymous with granulins A and B, respectively. Both the peptides and intact granulin protein regulate cell growth. However, different members of the granulin protein family may act as inhibitors, stimulators, or have dual actions on cell growth. Granulin family members are important in normal development, wound healing, and tumorigenesis. The CYFIP2 promoter contains a p53-responsive element that confers p53 binding as well as transcriptional activation of a heterologous reporter. So some genes which that been used in Ik-TSP classifier are very important, and it is obvious that the accuracy is higher than other classifiers. In addition, TCF3 is in acute myeloid leukemia pathway, so it is surely important for classifying between ALL and AML. CD33 is in Hematopoietic cell lineage pathway. Others are not directly correlated with Leukemia, but it is possible that they are correlated with genes in the pathway which is correlation with Leukemia.

For multi class gene expression dataset

We used the Ik-TSP classifier to distinguish specific genes among three subtypes of leukemia. Armstrong et al. (2002) identified specific genes involved in chromosomal translocation of the human acute leukemia known as the mixed-lineage leukemia (MLL). POU2AF1 is observed to be differentially expressed in the cells of patients with chronic lymphocytic leukemia. The NF2 gene provides instructions for the production of a protein called merlin, also known as schwannomin. This protein is made in the nervous system, particularly in specialized cells that wrap around and insulate nerves (Schwann cells). Merlin is believed to play a role in controlling cell shape, cell movement, and communication between cells. To carry out these tasks, merlin associates with the internal framework that supports the cell (the cytoskeleton). Merlin also functions as a tumor suppressor protein, which prevents cells from growing and dividing too fast or in an uncontrolled way. Somatic mutations in the NF2 gene are involved in the development of several types of tumors, both noncancerous (benign) and cancerous (malignant); hence it appears twice in the classification rule.

Mme encodes a common acute lymphocytic leukemia antigen that is an important cell surface marker in the diagnosis of human acute lymphocytic leukemia (ALL). This protein is presented on leukemic cells of pre-B phenotype, which represents 85% of cases of ALL. This protein is not restricted to leukemic cells, however, it is found on a variety of normal tissues. Cyclin G is a direct transcriptional target of the p53 tumor suppressor gene product and thus functions downstream of p53. GAK is an association partner of cyclin G. CSRP2 is a member of the CSRP family of genes, encoding a group of LIM domain proteins, which could be involved in regulatory processes and be important for development and cellular differentiation. CRP2 contains two copies of the cysteine-rich amino acid sequence motif (LIM) with putative zinc-binding activity, and may be involved in regulating ordered cell growth. BLNK encodes a cytoplasmic linker or adaptor protein that plays a critical role in B cell development. Deficiency in this protein has also been shown in some cases of pre-B acute lymphoblastic leukemia. For gene CD19, Lymphocytes proliferate and differentiate in response to various concentrations of different antigens. The ability of the B cell to respond in a specific, yet sensitive manner to the various antigens is achieved with the use of low-affinity antigen receptors. This gene encodes a cell surface molecule which assembles with the antigen receptor of B lymphocytes in order to decrease the threshold for antigen receptor-dependent stimulation. Furthermore, CHRNA7 is in calcium-signalling pathway, BLNK is in B cell receptor-signalling pathway, and CD19 is in hematopoietic cell lineage and B cell receptor-signalling pathway, which correlate with Leukemia. It can be seen that pathway-based methods have their advantages and promising.

Evidently, it is very difficult to classify the three types of Leukemia (Armstrong, 2004). They are acute lymphoblastic leukemia, acute myeloid leukemia and mixed-lineage leukemia. In this article the accuracy of Ik-TSP method is 100%, and it is better than any other method. Therefore, the improved method can be considered somewhat successful for the leukemia dataset.

Conclusion

In this paper, we presented the Ik-TSP to improve the original k-TSP algorithm. The proposed method avoids the shortcoming of the k-TSP method by using information entropy to select key genes from the dataset and integrating entropy with the k-TSP method. We also compared Ik-TSP classifiers with 5 different machine learning methods and the k-TSP method based on 3 different feature selection methods on 9 binary class gene expression datasets and 10 multi-class gene expression datasets involving human cancers. Experimental results showed that the Ik-TSP method had higher accuracy. In addition, the proposed method can effectively find genes that are important for distinguishing different cancer and cancer subtype.

ACKNOWLEDGEMENTS

This work was supported by the National Natural Science

Foundation of China (60673023, 10872077, 60873146 and 60673099), the National High-technology Development Project of China (2009AA02Z307), the European Commission (155776-EM-1-2009-1-IT-ERAMUNDUS-ECW-L12), the Science-Technology Development Project from Jilin Province (20080708); and “985” and “211” Project of Jilin University.

REFERENCES

Armstrong S, Staunton JE, Silverman LB, Pieters R, Boer MLD, Minder MD, Sallan SE, Lander ES, Golub TR, Korsmeyer SJ (2002). MLL translocations specify a distinct gene expression profile that distinguishes a unique leukaemia. Nat. Genet. 30: 41-47.

Armstrong SA (2004). Leukemia gene expression: MLL rearrangements in AML, ALL, Blood, 104: 3423-3424.

Acu E (2003). A comparison of filters and wrappers for feature selection in supervised classification, Proceedings of the Interface 2003 Computing Science, Statistics, p. 34.

Bø TH, Jonassen I (2002). New feature subset selection procedures for classification of expression profiles, Genome Biology 3, research0017.1-research00. pp.17.11.

Dietterich TG (2000). An experimental comparison of three methods for constructing ensembles of decision trees: bagging, boosting, and randomization, Machine Learning, 40: 139-157.

Dudoit S, Fridlyand J (2003). Classification in microarray experiments, In Speed TP (ed). Statistical Analysis of Gene Expression Microarray Data, Chapman, Hall/CRC, 93-158.

Geman D, d'Avignon C, Naiman DQ, Winslow RL (2004). Classifying gene expression profiles from pairwise mRNA comparisons, Statistical Applications Genet. Mol. Biol. 3, Article p. 19.

Gene Ontology Consortium (2004). The Gene Ontology (GO) database and informatics resource, Nucleic Acids Res. 32: 258-261.

Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, Lander ES (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring, Science, 286: 531-537.

Guyon I, Weston J, Barnhill S, Vapnik V (2002). Gene selection for cancer classification using support vector machines. Machine Learning, 46: 389-422.

Hedenfalk I, Duggan D, Chen Y, Radmacher M, Bittner M, Simon R Meltzer, P, Gusterson, B, Esteller, M, Kallioniemi OP, Wilfond B, Borg A Trent J, Raffeld M, Yakhini Z, Ben-Dor A, Dougherty E, Kononen J, Bubendorf L, Fehrle W, Pittaluga S, Gruvberger S, Loman N, Johannsson O, Olsson H, Sauter G (2001). Gene expression profiles in hereditary breast cancer. N. Engl. J. Med. 344: 539-548.

Helman P, Veroff R, Atlas S, Willman C (2004). A Bayesian network classification methodology for gene expression data. J. Computational Biol. 11: 581-615.

Khan J, Wei JS, Ringner M, Saal LH, Ladanyi M, Westermann F, Berthold F, Schwab M, Antonescu CR, Peterson C, Meltzer PS (2001). Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks, Nat. Med. 7: 673-679.

Kira K, Rendel L (1992). The Feature Selection Problem: Traditional Methods and a new algorithm, MIT Press, Cambridge.

Kononenko I, Simec E, Robnik-Sikonja M (1997). Overcoming the myopia of induction learning algorithms with Relieff. Appl. Intelligence, 7: 39-55.

Li T, Zhang CL, Ogihara M (2004). A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression. Bioinformatics, 20: 2429-2437.

Li YX, Ruan XG (2005). Cancer Subtype Recognition and Feature Selection with Gene Expression Profiles, Chinese J. Electronic Chinese, 33: 651-655.

Li Z, Bao, L, Huang YW, Sun ZR (2002). Cancer subtype discovery and information gene identification with gene expression profile, Acta Biophysica Sinica Chinese, 18: 413-417.

Liu CC, Chen WSE, Lin CC, Liu HC, Chen HY, Yang PC, Chang PC, Chen JJW (2006), Topology-based cancer classification and related pathway mining using microarray data, Nucleic Acids Res. 34: 4069-4080.

Mutis T, Verdijk R, Schrama E, Esendam B, Brand A, Goulmy E (1999). Feasibility of Immunotherapy of Relapsed Leukemia With *Ex Vivo*-Generated Cytotoxic T Lymphocytes Specific for Hematopoietic System-Restricted Minor Histocompatibility Antigens, Blood, 93: 2336-2341.

Shannon CE, Weaver W (2003). The Mathematical Theory of Communication, The University of Illinois Press, Urbana.

Tan AC, Gilbert D (2003). Ensemble machine learning on gene expression data for cancer classification, Appl. Bioinformatics, 2: S75-S83.

Tan AC, Naiman DQ, Xu L, Winslow RL, Geman D (2005). Online supplementary materials for simple decision rules for classifying human cancers from gene expression profiles, Available at https://jshare.johnshopkins.edu/atan6/public_html/KTSP.

Tan AC, Naiman DQ, Xu L, Winslow RL, Geman D (2005). Simple decision rules for classifying human cancers from gene expression profiles, Bioinformatics, 21: 3896-3904.

Wei JM, Wang SQ, Wang MY (2004). Novel approach to decision-tree construction, J. Adv. Computational Intelligence Intelligent Informatics, 8: 332-335.

Yang XJ (2004). The diverse superfamily of lysine acetyltransferases and their roles in leukemia and other diseases, Nucleic Acids Res. 32: 959-976.

Zhou XB, Liu KY, Wong STC (2004). Cancer classification and prediction using logistic regression with Bayesian gene selection. J. Biomed. Informatics, 37: 249-259.