A Comparative Study on Feature Selection and Classification Methods Using Gene Expression Profiles and Proteomic Patterns

Huiqing Liu huiqing@lit.a-star.edu.sg
Jinyan Li jinyan@lit.a-star.edu.sg
Limsoon Wong limsoon@lit.a-star.edu.sg
Laboratories for Information Technology, 21 Heng Mui Keng Terr, 119613 Singapore

Abstract

Feature selection plays an important role in classification. We present a comparative study on six feature selection heuristics by applying them to two sets of data. The first set of data are gene expression profiles from Acute Lymphoblastic Leukemia (ALL) patients. The second set of data are proteomic patterns from ovarian cancer patients. Based on features chosen by these methods, error rates of several classification algorithms were obtained for analysis. Our results demonstrate the importance of feature selection in accurately classifying new samples.

Keywords: proteomic patterns, gene expression profiles, feature selection, classification

1 Introduction

One of the important recent breakthroughs in experimental molecular biology is microarray technology. It allows scientists to monitor the expression of genes on a genomic scale. Such a technology increases the possibility of cancer classification and diagnosis at the gene expression level. However, many factors may affect the outcome of the analysis. One of them is the huge number of genes included in the original data. Some of them may be irrelevant to analysis. Thus, selecting discriminatory genes is critical to improving the accuracy and speed of prediction systems [1, 8, 7, 2].

This paper presents six feature selection heuristics based on entropy theory [6], $\mathcal{X}^2$-statistics, and t-statistics. After features are selected by these methods from the expression profiles, their effectiveness are investigated by comparing error rate of four traditional classification algorithms applied to only these selected features versus all features. The four classification algorithms are k-nearest neighbor (k-NN) [4], C4.5 [16], Naive Bayes (NB) [10], and Support Vector Machines (SVM) [3]. It will be seen that a great accuracy improvement can be achieved by the four classification algorithms if discriminatory features are first determined by the six feature selection methods.

Recently, we have developed a new classifier [18, 12] called PCL (Prediction by Collective Likelihood of emerging patterns [5]). The basic idea of PCL is to use a collection of multi-feature discriminators for classification. It is new because feature groups are used in PCL's input instead of the traditional use of individual features in the other classification algorithms. We also compare the performance of our PCL classifier with the four traditional classifiers to see whether PCL is competitive to the best of them.

The organization of this paper is as follows. The two data sets used in this study are described in Section 2. The feature selection methods and heuristics are outlined in Section 3. The five classification algorithms (including our PCL classifier) are briefly described in Section 4. Our experimental results are reported in Section 5. Our analysis on the results, including the significance of the feature selection methods, are presented in Section 6. Then we provide some discussions and future work in Section 7. Supplementary information (other tables and figures) can be found at `http://sdmc.lit.org.sg/GEDatasets`.

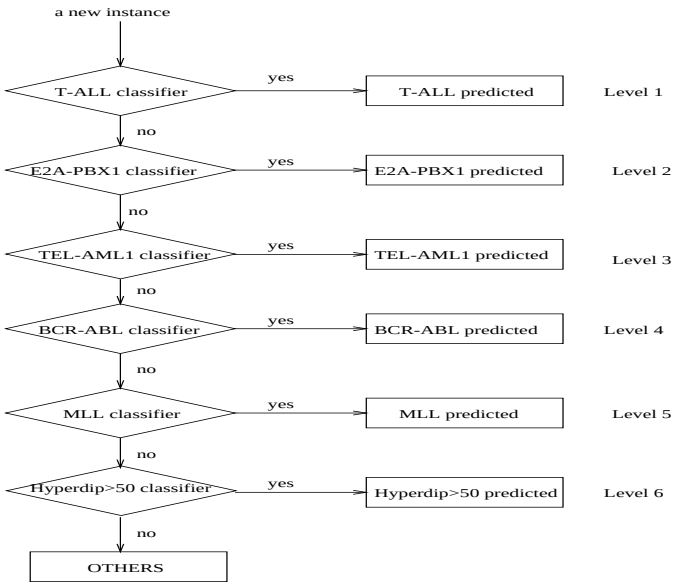

Figure 1: A tree-structure system for predicting more than six subtypes of ALL samples.

2 Two Sets of Data

Acute lymphoblastic leukemia (ALL). This is a collection of gene expression profiles of 327 ALL samples [18]. These profiles were obtained by hybridization on the Affymetrix U95A GeneChip containing probes for 12558 genes. These 327 samples contain all known ALL subtypes, including T-cell (T-ALL), E2A-PBX1, TEL-AML1, MLL, BCR-ABL, and hyperdiploid (Hyperdip>50). A tree-structured decision system (as shown in Figure 1), proposed by the medical doctors [18], is used to classify these samples. When a sample is given, rules are applied firstly for classifying whether it is a T-ALL or a sample of other subtypes. If it is classified as T-ALL, then the process is terminated. Otherwise, the process is moved to level 2 to see whether the sample can be classified as E2A-PBX1 or the remaining other subtypes. With similar reasoning, a decision process based this tree can be terminated at level 6 where the subtypes Hyperdip>50 and OTHERS are determined. The data was divided into a *training set* of 215 samples and a blind *testing set* of 112 samples by the medical doctors[18]. In accordance to Figure 1, we further subdivide the training and testing data into six pairs of subsets, one for each level of the tree. The data set names and their ingredients are given in Table 1. The raw gene expression data can be found at `http://www.stjuderesearch.org/data/ALL1/`. The processed data in the format .data and .names can be found at `http://sdmc.lit.org.sg/GEDatasets/Datasets#Leukemia`, that is commonly used by machine learning communities.

Ovarian cancer. The second group of data is a set of ovarian cancer samples that was first reported in [15]. The goal of this experiment is to identify proteomic patterns in serum that distinguish ovarian cancer from non-cancer. That study is significant to women who have a high risk of ovarian cancer due to family or personal history of cancer. The proteomic spectra were generated by mass spectroscopy and the raw data can be found at `http://clinicalproteomics.steem.com`. Since the web site is to display the most current data, it would be continually updated once new data sets and/or better models come out. Our experiments and results reported in this paper are based on the data set 6-19-02, which consists of 91 controls (non-cancer) and 162 ovarian cancers. In order to be consistent with [15], the relative amplitude of the intensity at each molecular mass/charge (M/Z) identity in the spectral data was normalized against the most intense and the least intense values in the data stream according to the formula: $NV = (V - Min)/(Max - Min)$, where NV is the normalized value, V the raw value, Min the minimum intensity and Max the maximum intensity. The normalization is done over all the 253 samples for all 15154 M/Z identities. After the normalization, each intensity value is

Table 1: Six pairs of training and testing data sets. The OTHERS1, OTHERS2, OTHERS3, OTHERS4, OTHERS5, and OTHERS classes consist of more than one subtypes of ALL samples. Columns 3 and 4 give the number of samples of the training and testing sets in the format of n_l vs n_r, where n_l (n_r) is the number of samples in the left-side (right-side) class. OTHERS1 = {E2A-PBX1, TEL-AML1, BCR-ABL, MLL, Hyperdip>50, OTHERS}; Similarly, OTHERS2 = {TEL-AML1, BCR-ABL, MLL, Hyperdip>50, OTHERS}; ...; OTHERS5 = {Hyperdip>50, OTHERS}.

Level	Paired datasets	Training set size	Testing set size
1	T-ALL vs OTHERS1	28 vs 187	15 vs 97
2	E2A-PBX1 vs OTHERS2	18 vs 169	9 vs 88
3	TEL-AML1 vs OTHERS3	52 vs 117	27 vs 61
4	BCR-ABL vs OTHERS4	9 vs 108	6 vs 55
5	MLL vs OTHERS5	14 vs 94	6 vs 49
6	Hyperdip>50 vs OTHERS	42 vs 52	22 vs 27

in the range between 0 and 1. The processed data in the format .data and .names can be found at `http://sdmc.lit.org.sg/GEDatasets/Datasets#OvarianCancer`.

3 Selecting Discriminatory Features

The number of features captured in the data is very large. We use an entropy-based [6], a $\mathcal{X}^2$-statistics, a correlation-based [9], a t-statistics, and a MIT correlation-based [8] feature selection method to filter out irrelevant features. The first three methods are commonly used in the machine learning community, while the last two are favored by earlier works on gene expression analysis. The $\mathcal{X}^2$-statistics and the correlation-based method are two refinements of the entropy method. The MIT correlation method can be considered as a variant of the t-statistics. We study them here and aim to compare their effectiveness in handling gene expression data. For simplicity, the word "features" is used in this paper to mean genes in ALL data and M/Z identities in ovarian cancer data.

3.1 Entropy-Based

The basic idea of this method [6] is to filter out those features whose expression distributions are relatively random. For the remaining features, this method can automatically find some cut points in these features' value ranges such that the resulting expression intervals of every feature can be maximally distinguished. If every expression interval induced by the cut points of a feature contains only the same class of samples, then this partitioning by the cut points of this feature has an entropy value of zero. This is an ideal case. The smaller a feature's entropy is, the more discriminatory it is. For a detailed description of the algorithm, please refer to [13, 11, 12] or `http://sdmc.lit.org.sg/gedm/Preprocessing.html`. We sort the values of entropy in an ascending order and consider those features with lowest entropy values.

3.2 The $\mathcal{X}^2$-Statistics and Correlation-Based Feature Selection Methods

The *Chi-Squared* ($\mathcal{X}^2$) method [14] and the *Correlation-based Feature Selection* (CFS) method [9] are built on the top of the entropy method.

The $\mathcal{X}^2$ method evaluates features individually by measuring their chi-squared statistic with respect to the classes. For a numeric attribute, the method first requires its range to be discretized into several intervals using, for example, the entropy-based discretization method. The $\mathcal{X}^2$ value of an attribute is defined as: $\mathcal{X}^2 = \sum_{i=1}^{m} \sum_{j=1}^{k} \frac{(A_{ij}-E_{ij})^2}{E_{ij}}$, where m is the number of intervals, k the number of classes,

A_{ij} the number of samples in the ith interval, jth class, R_i the number of samples in the ith interval, C_j the number of samples in the jth class, N the total number of samples, and E_{ij} the expected frequency of A_{ij} ($E_{ij} = R_i * C_j/N$).

After calculating the $\mathcal{X}^2$ value of all considered features, we can sort these values with the largest one at the first position, as the larger the $\mathcal{X}^2$ value, the more important the feature is. The degree of freedom of the above $\mathcal{X}^2$-statistics is $(m-1)\cdot(k-1)$. Since in most cases, there are only two intervals found for a feature (ie, $m = 2$), we use $k-1$ as the degree of freedom for the $\mathcal{X}^2$-statistics. It is equal to 1 in this paper. The critical $\mathcal{X}$ value for 1 degree of freedom at 5% significant level is 3.841 [17].

The CFS method is another approach to feature selection. Rather than scoring (and ranking) individual features, the method scores (and ranks) the worth of subsets of features. As the feature subset space is usually huge, CFS uses a best-first-search heuristic. This heuristic algorithm takes into account the usefulness of individual features for predicting the class along with the level of intercorrelation among them with the belief that "good feature subsets contain features highly correlated with the class, yet uncorrelated with each other". CFS first calculates a matrix of feature-class and feature-feature correlations from the training data. Then a score of a subset of features assigned by the heuristic is defined as: $Merit_S = \frac{k\overline{r_{cf}}}{\sqrt{k+k(k-1)\overline{r_{ff}}}}$, where $Merit_S$ is the heuristic merit of a feature subset S containing k features, $\overline{r_{cf}}$ is the average feature-class correlation, and $\overline{r_{ff}}$ is the average feature-feature intercorrelation.

Symmetrical uncertainties are used in CFS to estimate the degree of association between discrete features or between features and classes [9]. The formula below measures the intercorrelation between two features or the correlation between a feature X and a class Y which is in the range $[0, 1]$. $r_{xy} = 2.0 * [\frac{gain}{H(X)+H(Y)}]$, where $gain = H(X) + H(Y) - H(X, Y)$, is the information gain between features and classes, $H(X)$ is the entropy of the feature. CFS starts from the empty set of features and uses the best-first-search heuristic with a stopping criterion of 5 consecutive fully expanded non-improving subsets. The subset with the highest merit found during the search will be selected.

3.3 T-Statistics and MIT Correlation

An often used feature selection method is based on t-statistics. This method starts with a data set S consisting of m expression vectors: $X^i = (x_1^i, \cdots, x_n^i)$, where $1 \le i \le m$, m is the number of samples, and n is the number of features measured. Each sample is labeled with $Y \in \{+1, -1\}$ (for classes, such as T-ALL vs. OTHERS1). For each feature x_j, the mean μ_j^+ (resp. μ_j^-) and the standard deviation δ_j^+ (resp. δ_j^-) using only the samples labeled $+1$ (resp. -1) are calculated. Then a score $T(x_j)$ can be obtained by $T(x_j) = \frac{|\mu_j^+ - \mu_j^-|}{\sqrt{\frac{(\delta_j^+)^2}{n_+} + \frac{(\delta_j^-)^2}{n_-}}}$ where n_+ (resp. n_-) is the number of samples labeled as $+1$ (resp. -1). When making selection, we simply take those features with the highest scores as the most discriminatory features.

The score for each feature can be calculated by a slightly different formula as shown below. This method is called MIT correlation [8], which is also known as signal-to-noise statistic. The score is defined as: $MIT(x_j) = \frac{|\mu_j^+ - \mu_j^-|}{\delta_j^+ + \delta_j^-}$.

Observe that the entropy, $\mathcal{X}^2$, and CFS feature selection methods are resistant to data normalization. This means that each of these three methods will choose the same features regardless of whether the data is normalized by a pre-processing step, for example, by taking logarithms. This is not true for the t-statistics and MIT correlation methods, because the scores calculated by those formulas and the features' orderings can be different after the original data are processed with a normalization step.

3.4 Features Selection Heuristics Used in Our Experiments

We first consider the following six heuristics in our experiments:

- Select all the features recommended by CFS (all-CFS);
- Select top 20 features with the highest $\mathcal{X}^2$-statistics scores (top20-$\mathcal{X}^2$);
- Select top 20 features with the highest t-statistics scores (top20-t);
- Select top 20 features with the highest scores calculated by MIT formula (top20-mit);
- Select features having an entropy value less than 0.1 if these exist, or the 20 features with the lowest entropy values otherwise (Entropy); and
- Select all features which meet 5% significant level of $\mathcal{X}^2$-statistics (all-$\mathcal{X}^2$).

4 Classification Algorithms

After selecting the most discriminatory features, we apply k-NN, C4.5, NB, SVM, and PCL to obtain error rates on our testing samples. The classification results of these algorithms are then used to compare the effectiveness of various feature selection methods.

k-NN is a typical instance-based prediction model. By k-NN, the class label of a new testing sample is decided by the majority class of its k closest neighbors based on their Euclidean distance. In our experiments, k is set as 1.

C4.5 is a widely used decision tree based classifier. The implementation of C4.5 in this paper is based on its Revision 8, which was the last public version before it was commercialized. In our experiments, pruned trees and subtree raising techniques are used.

Naive Bayes (NB) is a probabilistic learner based on Bayes's rule. It is among the most practical approaches to certain types of learning problems.

SVMs are a kind of blend of linear modeling and instance-based learning. A SVM selects a small number of critical boundary samples from each class and builds a linear discriminant function that separates them as widely as possible. In the case that no linear separation is possible, the technique of "kernel" will be used to automatically inject the training samples into a higher-dimensional space, and to learn a separator in that space. The SVM used in this paper is a version that implements a sequential minimal optimization algorithm using polynomial kernels. Transforming the output of SVM into probabilities is conducted by a standard sigmoid function.

PCL is based on the concept of emerging patterns [5]. It needs to conduct a feature selection process before its model is established. Those selected features are then discretized. Then, the core knowledge patterns, our emerging patterns, are derived from the discretized training data. See the supplementary material of the work [18] for a full description of PCL. Note that an emerging pattern is a set of conditions often including several features, with which most of a class of samples' expression satisfy, but none of other class's samples satisfy. So, an emerging pattern can be considered as a multi-feature discriminator. The central spirit of PCL is to use top-ranked multi-feature discriminators to make a collective prediction. PCL uses feature groups, does not assume that features are independent; PCL can provide more than a mere prediction or a distance, but many interesting rules.

The main software package used in our experiments is *Weka* (Waikato Environment for Knowledge Analysis), developed at the University of Waikato in New Zealand. It is a powerful open source Java-based machine learning software package. It is publicly available online at `http://www.cs.waikato.ac.nz/ml/weka`. Also, we have implemented in-house programs, like the discovery of emerging patterns, PCL, the entropy feature selection, t-statistics and MIT correlation feature selection methods.

Table 2: A summary of the number of misclassified testing samples for all the five classifiers for ALL data.

Selection	Total misclassifications of six levels					Average
heuristics	SVM	NB	k-NN	C4.5	PCL	
all-CFS	6	12	7	12	-	9.25
top20-$\mathcal{X}^2$	6	8	7	14	4	8.75
top20-t	10	19	10	18	5	14.25
top20-mit	7	7	9	14	4	9.25
Entropy	5	7	4	14	5	7.5
all-$\mathcal{X}^2$	3	13	7	15	-	9.5
all genes	23	63	23	26	-	33.75

5 Experimental Results

5.1 Results of ALL Data

Recall that we have six levels of testing samples. For each level and for each feature selection heuristic, we apply the classification algorithms to obtain our results. We note that every feature selection heuristic and classification model is built on the training data only. We mostly use the *error rate* of testing samples to illustrate our results. The format of the error rate $x : y$ means that x number of samples in the left-side class (e.g., T-ALL, E2A-PBX1 and etc.) are misclassified, and y number of samples in the right-side class are misclassified (please refer to the data description in Table 1).

Results on the original data without feature selection. Table S1 in our Supplementary Information reports the error rates of the four state-of-the-art classification algorithms on all six levels of testing samples. In total, SVM, NB, k-NN, and C4.5 made 23, 63, 23, and 26 misclassifications respectively. These error rates are not acceptable for diagnostic purposes.

Results on data consisting of only selected genes. For the training data at Level 1, the CFS method selected only 1 gene (38319_*at*) from the total of 12558 genes. There were 13 genes with entropy values less than 0.10. The number of genes selected by all-$\mathcal{X}^2$ was 1309. SVM, NB, k-NN, and PCL achieved 100% prediction accuracy on this level's testing samples for all the six groups of genes selected via the six selection heuristics. However, C4.5 misclassified a same particular testing sample (from the "OTHERS1" class) for all the testing data based on the six groups of selected genes. At level 2, CFS also selected only 1 gene (33355_*at*) as the most discriminatory gene. There were 8 genes with entropy values less than 0.10. The number of genes selected by all-$\mathcal{X}^2$ was 827. All five classification algorithms achieved 100% accuracy on this level testing data no matter which feature selection heuristic was used. After feature selection, the classification algorithms performed excellently on the testing data of Levels 1 and 2. For the remaining four levels of data, the results were not that perfect but were still very good. The error rates of Level 3 to Level 6's testing data varied for different learning algorithms and for different feature selection heuristics. Tables S2, S3, S4 and S5 in our Supplementary Information report these error rates. (A "-" sign means that the results are not available at the deadline of submission of the paper.) We use Table 2 to summarize total misclassifications of all six levels for every classifier under different feature selection heuristics. For a fixed feature selection method, the average misclassifications over the four traditional classifiers are also reported in the last column of Table 2.

5.2 Results of Ovarian Cancer Data

For this data set, we use our six feature selection heuristics to choose important M/Z identities and apply SVM, NB, k-NN, and C4.5 to obtain error rates of running 10-fold cross validation on all 253

Table 3: Error rates on ovarian cancer data set with or w/o feature selection.

Selection heuristics	# of selected feature	SVM	NB	k-NN	C4.5	Average
all-CFS	17	0:0	0:3	0:0	4:3	2.5
top20-$\mathcal{X}^2$	20	2:3	2:2	2:3	5:4	5.75
top20-t	20	2:2	2:3	2:4	5:3	5.75
top20-mit	20	2:3	3:2	2:3	5:4	6
Entropy	20	2:3	2:2	2:3	5:4	5.75
all-$\mathcal{X}^2$	6136	0:0	10:6	2:7	5:4	8.5
all features	15154	0:0	17:2	6:9	4:5	10.75

samples. Table 3 shows the error rates of the four traditional classification algorithms running 10-fold cross validation on all 253 samples under different feature selection heuristics. The last row gives the results on original data with 15154 M/Z identities. For this data set, the error rate format $x : y$ means that x numbers of ovarian cancer samples and y numbers of non-cancer samples are misclassified. In this test, CFS selected only 17 features, but performed very well. With these 17 M/Z values, both SVM and k-NN achieved 100% classification accuracy. Besides, SVM got 100% accuracy with all-$\mathcal{X}^2$ selected M/Z values and PCL misclassified only one sample with Entropy selected features (data not shown). On the other hand, although the results on the full set of 15154 M/Z identities is not very bad, the response speed of each algorithm is very slow.

In [15], an approach was designed to use genetic algorithm and self-organizing cluster analysis to locate some key M/Z values that best distinguish cancer from non-cancer. On their web site, there is a group of best M/Z values provided. It contains seven key M/Z values: 2760.6685, 19643.409, 465.56916, 6631.7043, 14051.976, 435.4652 and 3497.5508. However, among these seven M/Z values, we found 3 of them are having 0 $\mathcal{X}^2$ value. They are: 2760.6685, 19643.409 and 6631.7043. If we remove these 3 from the group, there are only 4 identities left. The error rates of 10-fold cross validation using these seven or four key M/Z values are indeed very close by SVM, NB, k-NN and C4.5 (data not shown). Remarkably, SVM achieved 100% accuracy with these 4 M/Z values. Furthermore, these 4 key M/Z values can be regarded as another example of best key M/Z values combination and $\mathcal{X}^2$-statistics is successfully used to narrow down the features selected by other approach.

6 Comparison and Analysis

First of all, in most of cases, the accuracy performance of the four classifiers were greatly improved after features are selected by our proposed heuristics. For example, at every level's testing samples of ALL data, all the four traditional classifiers have improved their poor performance or maintained their excellent accuracy after features are selected by most of our proposed heuristics. This is particularly true for level 6. Next, we highlight other interesting points as follows:

- Of the six proposed selection heuristics, overall speaking, for ALL data set, the entropy one appeared to be the best. Under this strategy, NB made its smallest number of errors of 7; k-NN made its smallest number of errors of 4. Both SVM and C4.5 made their the second smallest number of errors of 5 and 14 respectively. This can be also seen from the sixth column of Table 2 where the average number of errors per classifier under a specific feature selection heuristic is summarized. Note that the average number under the entropy scheme was 7.5, the smallest.

- The relatively new selection method, CFS, also demonstrated its feasibility on gene or protein expression profiles. In ovarian cancer data test, with the only 17 features selected by CFS, both

SVM and k-NN achieved 100% accuracy. NB and C4.5 also got best results under this heuristic. An advantage of CFS over other methods is that it can automatically determine the number of discriminatory features. Sometimes, CFS can return very small number of features, for example, only one gene at level 1 and level 2 of ALL data. However, the performance of the classifiers based on those small number of genes were very good.

- Of the four traditional classification algorithms, SVM appeared to perform best even on the original intact data. SVM contains non-linear kernel functions so that non-linear mappings and decisions can be easily achieved. C4.5 can provide small-sized elegant decision trees to classify testing samples. Its accuracy was not excellent, though its performance on the *training* data closely approached to the perfect level. Interestingly, the simplest k-NN classifier provided us a very good accuracy. This indicates that after feature selection the expression data can be well clustered according to distance. The Naive Bayes classifier has an important assumption that features should be independent. So, it's not surprising that its performance was not good because many feature groups interact closely.

- For our new PCL classifier, its performance was competitive to SVM. An advantage of PCL over SVM is that PCL can provide users many high-level and comprehensible rules.

7 Discussion and Future Work

In this section, we discuss three questions. The first question is that if 20 features are randomly selected from the total features, how far the resulting accuracy is from the accuracy based on the top 20 features selected by our proposed heuristics? We address this question to see if our selected 20 features are collectively outstanding.

Recall that the threshold of 20 used to cut off top ranked features is an arbitrary number, though it is based on our experience. So the second question is to study the accuracy trend when varying the number of selected top ranked features.

Our third question is to find how common are two groups of features selected by different heuristics. Then we can understand more about which features are playing a discrimination role.

Misclassifications if 20 features are randomly selected

Our experimental procedure is (1) randomly select a set of 20 features, and extract corresponding training and testing data from the original data. (2) apply the four traditional classification algorithms on the dimension-reduced data to get the number of misclassifications on the testing data. These two steps are repeated 100 times. Then the average and the minimum number of misclassified testing samples over the 100 experiments are reported, as shown in Table 4 for ALL data. The result of this test on ovarian cancer data for 10-fold cross validation is: average (minimum) number of misclassified samples for SVM is 79 (49), NB 94 (70), k-NN 81 (30) and C4.5 76 (31). We found that: the accuracy of the four traditional classifiers became very poor if the 20 features are randomly selected from the original data. All are much worse than that with the features selected by our heuristics. In most cases, for a specific classifier, its worst accuracy by our proposed six feature selection heuristics is still better than the best one when using randomly picked-up features. Therefore, it can be seen that the 20 features selected by our heuristics are much more discriminatory than those randomly selected features.

Accuracy trends when varying the number of selected features

For ALL data, we used these numbers, 5, 10, 20, 50, 100, 150, 200, 250 and 300 to select top-ranked features by Entropy. The error rates (number of misclassified samples) of SVM and C4.5 are reported

Table 4: The average (minimum) number of misclassified testing samples over 100 experiments each with a set of 20 randomly selected genes at each level.

Level	Ave.(min.) # of misclassifications			
	SVM	NB	k-NN	C4.5
1	13.4 (1)	28.4 (6)	17.6 (4)	13.3 (2)
2	8.7 (3)	33.8 (2)	13.6 (6)	9.5 (4)
3	25.2 (8)	31.8 (13)	30.4 (14)	25.6 (11)
4	6.01 (6)	20.8 (5)	9.4 (5)	6.6 (4)
5	6.3 (3)	16.3 (4)	9.4 (3)	7.4 (3)
6	16.5 (4)	18.6 (6)	19.0 (5)	17.2 (5)

in Figure S1 and S2 in our Supplementary Information based on this series of gene groups for the 6 levels data. In general, we found that there was no regular rule to determine an optimal number of features to get the best accuracy even for a specific classifier. We believe that the optimal number of the most discriminatory features may change from data to data, may depend on classification algorithms, and also may vary from different feature selection methods.

The number of 20 is used in some of our experiments as we had three considerations: (i) Medical doctors and biologists like a small number of features to separate two classes of cells. Manually examining a large amount of features is tedious and sometimes impossible; (ii) The decision speed should be fast. Some classifiers would need a long time to complete their learning phase if thousands of features are selected. (iii) For a classification problem where only two classes are involved, a small number of most discriminatory features could be used to distinguish the two classes well. Otherwise, even with more number of features, the distinction would not be necessarily become better.

How common are those features selected by two different heuristics?

Figure S3 and Table S6 in our Supplementary Information reports the number of overlapping features between two groups of features selected by two different heuristics for ALL data and ovarian cancer data. Observe that, the numbers and their distributions changes from data to data. We found that the top20-$\mathcal{X}^2$ and Entropy heuristics agreed on the most discriminatory features at the most degree. On average across the four levels of ALL data, Table S7 in our Supplementary Information, about 75% of the two groups of features ranked by them are identical. For ovarian cancer data, the features selected by top20-$\mathcal{X}^2$ and Entropy heuristics are same.

We also applied the four traditional classification algorithms to the ALL data using the common features selected by top20-$\mathcal{X}^2$ and Entropy. In most cases, we got slightly better or equal results from level 3 to 6's testing samples.

Future work

As shown in our experiments, feature selection is very important for classifying gene and protein expression data. For the future work, We will widen our scope to consider more feature selection and classification algorithms such as boosting, genetic algorithm, evolutionary algorithm, and neural networks. So that we can find an optimal approach to determining discriminatory features. To find common features from different feature selection methods is another interesting problem. We may also consider the union of features selected by different methods.

References

[1] Alon, U. *et al.*, Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays, *Proc. Natl. Acad. Sci. USA*, 96:6745–6750, 1999.

[2] Ben-Dor, A. *et al.*, Tissue classification with gene expression profiles, *Journal of Computational Biology*, 7:559–583, 2000.

[3] Burges, C., A tutorial on support vector machines for pattern recognition, *Data Mining and Knowledge Discovery*, 2:121–167, 1998.

[4] Cover, T. M. and Hart, P. E., Nearest neighbor pattern classification, *IEEE Transactions on Information Theory*, 13:21–27, 1967.

[5] Dong, G. and Li, J., Efficient mining of emerging patterns: Discovering trends and differences, *Proc. Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 43–52, 1999.

[6] Fayyad, U. and Irani, K., Multi-interval discretization of continuous-valued attributes for classification learning, *Proc. 13th International Joint Conference on Artificial Intelligence*, 1022–1029, 1993.

[7] Furey, T.S. *et al.*, Support vector machine classification and validation of cancer tissue samples using microarray expression data, *Bioinformatics*, 16:906–914, 2000.

[8] Golub, T. R. *et al.*, Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring, *Science*, 286, 531–537, 1999.

[9] Hall, M.A., Correlation-based feature selection machine learning, Ph.D. Thesis, Department of Computer Science, University of Waikato, Hamilton, New Zealand, 1998.

[10] Langley, P., Iba, W., and Thompson, K., An analysis of Bayesian classifier, *Proc. Tenth National Conference on Artificial Intelligence*, 223–228, 1992.

[11] Li, J. and Wong, L., Identifying good diagnostic genes or genes groups from gene expression data by using the concept of emerging patterns., *Bioinformatics*, 18:725–734, 2002.

[12] Li, J. *et al.*, Simple rules underlying gene expression profiles of more than six subtypes of acute lymphoblastic leukemia (ALL) patients, *Bioinformatics*, in press.

[13] Li, J. and Wong, L., Emerging patterns and gene expression data, *Genome Informatics*, 12:3–13, 2001.

[14] Liu, H. and Setiono, R., Chi2: Feature selection and discretization of numeric attributes, *Proc. IEEE 7th International Conference on Tools with Artificial Intelligence*, 338–391, 1995.

[15] Petricon, E.F. *et al.*, Use of proteomic patterns in serum to identify ovarian cancer, *The Lancet*, 359:572–577, 2002.

[16] Quinlan, J.R., *C4.5: Programs for Machine Learning*, Morgan Kaufmann, 1993.

[17] Sandy, R., *Statistics for Business and Economics*, McGrawHill, 1989.

[18] Yeoh, E.-J. *et al.*, Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling, *Cancer Cell*, 1:133–143, 2002.