Feature Space Transformation for Better Understanding Biological and Medical Classifications

Jinyan Li and Hwee-Leng Ong

Institute for Infocomm Research
21 Heng Mui Keng Terrace
Singapore 119613
Email: {jinyan, hweeleng@i2r.a-star.edu.sg}

Recently published gene expression profiles and proteomic mass/charge ratios are extremely high-dimensional data. Though support vector machines can well learn the inner relationship of the data for classification, the non-linear kernel functions pose an obstacle to explain the prediction reasons to non-specialists. In this paper, we study the problem of feature space transformation for easy interpretability of classification results. Each new feature is a combination of multiple original features provided that the new feature captures a large percentage of one class of data, but sharply discriminates the data in the other class. Under the description of new features, training or test data are clearly class-separable. We also discuss a more sophisticated rule-based method, called PCL, for classification. PCL provides easily explainable classification scores for us to better understand the predictions and the test data themselves. Visualization is also used to enhance the understanding of the classifier output. We use rich examples to demonstrate our main points.

ACM Classification: 1.2 (Artificial Intelligence), J.3 (Life and Medical Sciences)

1. INTRODUCTION

Classification, or making decisions on test data, is an important research topic in computer-based diagnostic medicine. The process is as follows: the classification algorithms (the classifiers) learn a mapping between the attribute values of training data and class labels, and then classify each test instance (sample) as one of the previously known classes. For example, the widely-used support vector machines (SVMs) (Burges, 1998) use learned non-linear kernel functions to transform weighted sum of attributes' values, and then take a step-function to categorize the input as a prediction. For another example, the *k*-nearest neighbour (*k*-NN) (Cover and Hart, 1967) chooses distance functions to classify test data: the class of the nearest training instance is predicted as the class of a test sample. In diagnostic medicine, the reasons provided by the classifiers can sometimes be as important as the prediction outcome themselves because medical doctors need the

A previous version of this paper appeared at First Asia-Pacific Bioinformatics Conference, Adelaide, Australia.
Correspondence Author: Jinyan Li.

Manuscript received: 3 May 2003
Communicating Editor: Phoebe Chen

explanations of the diagnosis. The "reasons" provided by the SVMs are the learned settings of the kernel functions. The "reasons" provided by *k*-NN are the types of distance measurements.

These reasons either are difficult to explain to non-specialists (as in SVMs), or are too simple to provide any more useful information (as in *k*-nearest neighbour). So, in this paper, we address the problem of outcome explainability by:

1. **Feature space transformation.** Based on the original features (attributes), we generate new features that have more powerful difference description strength. Under the new feature space, transformed *training* data are tidy and apparently class-separable. Furthermore, transformed *test* data are also visually decidable.
2. **Rule-based classification.** We do not use non-linear kernel functions as a basis to construct a classifier. Instead, we use discrimination rules (patterns) as the core idea of our PCL classifier (Li and Wong, 2002a; Li, Liu, Downing, Yeoh and Wong, 2003; Yeoh *et al*, 2002). Compared to non-linear functions, rules are easily readable and understandable to common users. Another important factor is that the rule-based classifier is highly accurate; its performance has been shown to be competitive to the best performance of other classifiers.
3. **Visualization of decision results.** We use numerical scores to illustrate the evidence of the decisions, giving the users (medical doctors) a strong sense of confidence on why such a decision was made by the computer. We use a 2-dimensional display to visualize the reasons behind our classifier. The visualization may also help us to understand the mechanism of the disease, and to reveal subtle structure and relationships in the data that are not apparent from the classifier.

All of these are aimed to show that the computer's prediction is accurate, reasonable, and explainable.

1.1 Background

The central concept used in the paper is called emerging patterns (Dong and Li, 1999). In its simplest form, an emerging pattern (EP) is a group of conditions with which most or some samples of a class (say positive class) satisfy, but none of another class (say negative class) satisfy. The following is an example of EP, as discovered from a gene expression profiling of the prostate disease (Singh *et al*, 2002).

{gene(37720_at) > 215, gene(38028_at) <= 12}

This EP contains two conditions: (1) the expression of gene 37720_at is > 215; and (2) the expression of gene 38028_at is <= 12. This pattern is an EP since about 73% of prostate disease cells satisfy the two conditions, but no normal cells satisfy. Note that this EP groups multiple features, namely the two genes, on different constraints (the expression ranges). Many other similar patterns are also found to be dominant in one class but absent in another class. In fact, EPs are patterns that show sharp contrasts between classes. Having this property, we can use EPs as new features to re-organize and re-describe the original data, because by definition even a single new feature can sharply differentiate the two classes.

In this paper, we will take gene expression profiles (Yeoh *et al*, 2002; Singh *et al*, 2002) and proteomics data (Petricoin *et al*, 2002) as examples to illustrate our main points. Gene expression and proteomic profiling are two key technologies in post-genome medical diagnosis. (i) Microarray DNA chip technology has become a standard tool for studying patterns and dynamics of the genetic mechanisms of living cells. Gene expression profiling not only provides an attractive alternative to

current standard techniques such as histology, genotyping, and immunostaining for tumour classification, but also provides valuable insights into the molecular characteristics of specific cancer types. (ii) Proteomics technology can identify low molecular weight molecules in a high-throughput, non-biased discovery approach using patient serum, plasma, urine, or other secretions. Distinctive protein signatures can be produced to show high discriminatory potential, a desirable characteristics for medical diagnosis.

1.2 Related work and paper organization

Our PCL classifier (Li *et al*, 2003; Yeoh *et al*, 2002), or *prediction by collective likelihood from emerging patterns*, integrates the power of multiple top-ranked emerging patterns to form compact classification scores. With reference to the scores rising from training data, the classification scores of test data obtained by PCL can give a strong confidence indication on the decisions.

This paper reports PCL's performance on several data sets that are not studied in our previous works. Together with the previous results, we can see that PCL is a highly-accurate classifier indeed. For the first time, we use a score decomposition approach and a visualization tool to understand the PCL classifier and to explain the reasons behind the predictions. We also view the decomposition components as a new type of feature transformation.

In the next section, we discuss how to transform the original features into new features. Each new feature is an emerging pattern, grouping several genes together with some constraints on their expression range. We present the transformed representations of both training and test data in Section 3. In Section 4, we review the PCL classifier and report its high performance on several data sets. In Section 5, we analyse the classification scores derived by PCL, and then explain classification reasons behind PCL using a score decomposition method and a visualization tool. Section 6 concludes this paper.

2. FEATURE SPACE TRANSFORMATION: GENERATING NEW FEATURES TO REPLACE OLD ONES

We begin with a traditional representation of gene expression or proteomic data. Under this typical representation, it is hard for us to make visual predictions. Then we present steps to discover emerging patterns and to use these EPs to replace old features.

Sample	$gene_1$	$gene_2$	…	$gene_m$	Class
1	1002.3	123.7	…	10.5	normal
2	30.5	2543.1	…	21.0	abnormal
⋮	⋮	⋮	…	⋮	⋮
n	780.3	500.1	…	87.6	normal

Table 1: A gene expression data consist of *n* cell samples, each described by the expression levels of *m* genes. The samples are categorized into the normal or abnormal class shown in the last column.

2.1 A typical representation of gene expression profiles: A relational scheme

In the machine learning and bioinformatics communities, a gene expression profiling is commonly represented by a relational table, consisting of *n* tuples. Each tuple is described by the expression levels of *m* genes. Table 1 gives the structure of the representation.

Observe that each feature in such representations describes an expression range of a specific single gene. These features do not capture any interaction information of gene groups. However

interaction of gene groups are more interesting and more important to be used to differentiate normal and tumour cells because genes do not function in isolation. Emerging patterns can capture interesting information about the interaction.

2.2 Discover emerging patterns and use them as new features

Gene expression profiles or proteomic data usually consist of over ten thousand features, and most of them are not useful for classification. For fast discovery of significant emerging patterns, we need to remove those irrelevant features (genes) before conducting the discovery. We describe the steps as follows:

1. Using the entropy method (Fayyad and Irani, 1993), rank individually the features in terms of their power to distinguish two classes;
2. Select and discretize the top-ranked features. Usually, we select 20 top features;
3. Discover emerging patterns from the discretized data by a naive method (Li *et al*, 2003) or border-based algorithms (Li and Wong, 2002b; Dong and Li, 1999; Li, Ramamohanarao and Dong, 2000).

The entropy-based discretization method (Fayyad and Irani, 1993) can automatically remove about 90–95% of the whole feature space as those features exhibit random expression distributions. It can also automatically detect "ideally discriminatory genes" and "sub-optimal discriminatory genes" that contain clear boundaries separating two classes of cells. The role of discretization is to find a best cut point to partition an expression range of a gene such that every interval contains a same class of points as many as possible. The selection of top-ranked features can increase the possibility of producing significant emerging patterns.

Frequency is an important property of emerging patterns. The *frequency* of emerging patterns says the percentage of a class data satisfying the conditions contained in the pattern. The larger an emerging pattern's frequency is, the more important it is. Table 2 shows examples of emerging patterns discovered from a prostate disease data set (Singh *et al*, 2002) that consists of 52 Tumour samples and 50 Normal samples. As seen from Table 2, two characteristics of emerging patterns are: (a) they group several genes together; (b) they occur frequently in one class but are absent in the other class. So, we can view each emerging pattern as a multi-variate property of a class.

Therefore we can use emerging patterns to replace the original features, and then to re-describe the original data in a "tidy" way as described in the next section in detail. The feature transformation is outlined in Figure 1.

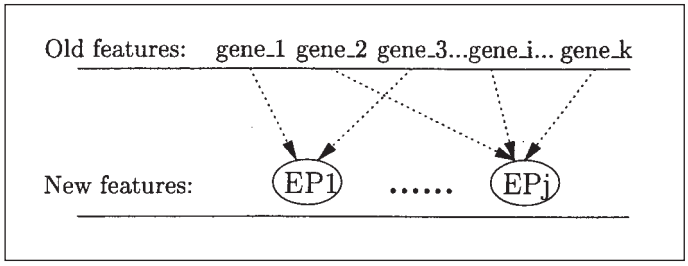

Figure 1: The feature space transformation. An important change is that a new feature is a combination of multiple original features. The grouping of multiple genes reflects a natural way to analyse gene expression profiles.

EPs	Occurrence (Frequency) in Tumour class	Occurrence (Frequency) in Normal class
{9,14}	38(73.08%)	0
{4,9}	38(73.08%)	0
{6,9}	38(73.08%)	0
{9,19}	37(71.15%)	0
{9,16}	37(71.15%)	0
{7,21}	0	36 (72.00%)
{7,11}	0	35 (70.00%)
{1,7}	0	32 (64.00%)
{5,7}	0	32 (64.00%)
{7,17}	0	31 (62.00%)

Table 2: A partial list of top-ranked EPs discovered from the prostate disease dataset when top 10 features are selected. The reference numbers contained in the patterns each represent a condition. For example, the reference number 9 in the first EP stands for the condition "the expression of gene 37720_at > 215", and similarly for other numbers. We did not include the meaning of all reference numbers.

gene_1	gene_2	gene_3	gene_4	gene_5	gene_6	gene_7	gene_8	gene_9	gene_10	class
16624.7	16517.9	7309.4	13148	5028.9	8822.9	4056.3	0	4801.9	24918.3	MLL
0	13992.2	45000	0	0	6141.3	0	0	0	32130.3	MLL
15099.1	6114.4	45000	15799.4	3156.8	6651.5	2784.4	0	5049.2	40961.4	MLL
18606.7	0	45000	11161.8	7184.4	6312.5	4500.9	0	0	33639.9	MLL
8104.9	9300.6	39102.5	0	2698.9	7573.4	0	0	3490.6	23798.5	MLL
9553	0	0	0	7185.8	4016.9	4870.6	45000	13718.1	6901.4	OTHERS
11993.5	0	10945.7	0	8678.8	2554.6	4282.2	45000	4434.3	34627.9	OTHERS
7892.5	0	45000	0	5537.6	0	6326.7	7589.8	4673.3	28125.6	OTHERS
11873.8	0	6163.1	0	6403.6	3790.2	7185.3	45000	7023.7	13619	OTHERS
8048	0	3319.6	7870.6	0	5078.2	8185.3	45000	0	8625.6	OTHERS

Table 3: Ten samples (5 MLL and 5 OTHERS) under the description of 10 top-ranked original features. The representation shows difficult visual differentiation rules to separate the two classes.

EPs	Occurrence (Frequency) in MLL class	Occurrence (Frequency) in OTHERS class
{9,13,15,21}	14(100.00%)	0
{2,15,21}	11(78.57%)	0
{12,13,21}	11(78.57%)	0
{9,12}	11(78.57%)	0
{2,13}	11(78.57%)	0
{1,3}	0	193 (96.02%)
{1,5,7}	0	192 (95.52%)
{1,11}	0	186 (92.54%)
{5,7,11}	0	185 (92.04%)
{3,7,11}	0	183 (91.04%)

Table 4: Five top-ranked EPs in the MLL class and five top-ranked EPs in the OTHERS class. We use these 10 EPs as new features to re-describe the data.

EP_{p1}	EP_{p2}	EP_{p3}	EP_{p4}	EP_{p5}	EP_{n1}	EP_{n2}	EP_{n3}	EP_{n4}	EP_{n5}	class
new features for MLL					new features for OTHERS					
1	1	1	1	1	0	0	0	0	0	MLL
1	0	1	1	0	0	0	0	0	0	MLL
1	1	1	1	1	0	0	0	0	0	MLL
1	1	1	1	1	0	0	0	0	0	MLL
1	0	1	1	0	0	0	0	0	0	MLL
0	0	0	0	0	1	1	1	1	1	OTHERS
0	0	0	0	0	1	1	1	1	1	OTHERS
0	0	0	0	0	1	0	1	0	1	OTHERS
0	0	0	0	0	1	1	1	1	1	OTHERS
0	0	0	0	0	1	1	1	1	1	OTHERS

Table 5: Top-ranked emerging patterns are viewed as new features to transform the original data. Under the new feature representation, training data can have a clear distinction between the two classes. See in the table the two sub-matrixes consisting of pure 0's.]

gene_1	gene_2	gene_3	gene_4	gene_5	gene_6	gene_7	gene_8	gene_9	gene_10	class
15761.2	8013	95176.4	11851.5	4646.9	4264.3	2943.2	0	2655.9	27961.2	unknown
13394.2	4886.1	82066.2	17983.6	5433.1	7070.4	3066.4	0	3448.7	38251.5	unknown
0	2932.5	0	0	12567.6	4873.3	7935.3	57555.3	6883.5	20223.5	unknown
20350.8	3771.1	44472.4	10809.9	5849.2	7869.9	4157.2	0	3300	43944.4	unknown
2932.1	0	0	0	5877.2	0	4735	0	7034.8	36802.3	unknown
16772.7	19834.4	51733.6	9256.1	0	6881.4	6614.3	0	0	28302.4	unknown
9114.8	0	3353.2	0	6641.4	2450.5	0	4731	5881.1	63644.6	unknown
20768	7634.1	65195.4	15401.9	4534	7739	4171.9	0	4160	29143.7	unknown
13157.8	0	12197.7	0	6872.4	4790.9	9781.4	0	11491.5	48053	unknown
8143.6	0	7604.1	0	5650.2	2148.1	5969	0	6715.5	49490.7	unknown

Table 6: Ten test samples under the description of 10 original features (genes). Their class prediction is visually difficult.

EP_{p1}	EP_{p2}	EP_{p3}	EP_{p4}	EP_{p5}	EP_{n1}	EP_{n2}	EP_{n3}	EP_{n4}	EP_{n5}	predicted class
new features for MLL					new features for OTHERS					
1	1	0	0	1	0	0	0	0	0	MLL
1	0	1	1	0	1	0	0	0	0	MLL
0	0	0	0	0	1	1	1	1	1	OTHERS
1	1	1	1	1	0	0	0	0	0	MLL
1	0	0	0	0	1	1	1	1	1	OTHERS
0	1	0	1	0	0	0	0	0	0	MLL
1	0	0	0	0	1	1	1	1	1	OTHERS
1	1	1	1	1	0	0	0	0	0	MLL
0	0	0	0	0	1	1	1	1	1	OTHERS
0	0	0	0	0	1	1	1	1	1	OTHERS

Table 7: After feature and data transformation, the class prediction of the 10 test samples becomes much easier.

3. TRANSFORMATION OF THE ORIGINAL DATA AND VISUALIZATION OF THE TRANSFORMED DATA

We use another data set as an example to demonstrate the procedure of data transformation and visualization. The data set consist of independent real test samples. The data is about subtype classification of childhood leukemia (Yeoh *et al*, 2002). The whole data consists of gene expression profiles of 327 cell samples. There are 14 training and 6 test samples of the subtype MLL (one of previously well studied subtypes), and 199 training and 106 test samples of all other subtypes. Table 3 shows a traditional, relational representation of 5 MLL training samples and 5 training samples of the other subtypes under the description of 10 top-ranked original features which are selected by the entropy method. The names of the 10 features are 34306_at, 36777_at, 33412_at, 657_at, 32207_at, 33847_s_at, 34337_s_at, 1389_at, 34861_at, and 40518_at. We denote them $gene_1, \cdots, gene_{10}$ respectively. (In the original data, there are 12558 features.)

Observe that if you do not look at the class labels, it is hard for you to tell us which sample belongs to MLL or which sample belongs to OTHERS. However, under our new features' representation, this problem becomes much easier to answer.

3.1 Representing the original data under new feature space

We discovered a total of 33 emerging patterns in the MLL class, and 24 emerging patterns in the OTHERS class. Table 4 shows the 10 most frequent EPs.

Basically, the transformation is conducted as follows:

1. Denote the 10 EPs (sequentially from the top of Table 4 as $EP_{p1}, \cdots, EP_{p5}, EP_{n1}, \cdots, EP_{n5}$;
2. Use $EP_{p1}, \cdots, EP_{p5}, EP_{n1}, \cdots, EP_{n5}$ as new features, the value of each feature is categorical, taking either 1 or 0. The value 1 of a feature means that the feature (the EP) is contained in a sample. Otherwise, it is the value 0;
3. Given a sample, test the 10 features (EPs) to see which of them are contained in the sample. Then use the test results (1 or 0) to represent the sample.

Consequently, the traditional representation of the 5 MLL training samples and the 5 training samples of the other subtypes, shown in Table 3, are transformed as in Table 5. The transformed second MLL sample says that this sample contains only EP_{p1}, EP_{p3}, and EP_{p4} but not other new features.

By definition, an emerging pattern occurs in only a class of training data, but does not occur in the other class of training data. So, for the 5 MLL training samples, the values of the features $EP_{n1}, \cdots, EP_{n5}$ must be 0; the values of the features $EP_{p1}, \cdots, EP_{p5}$ may be 1 or 0. Similarly, for the 5 OTHERS training samples, the values of the features $EP_{n1}, \cdots, EP_{n5}$ may be 1 or 0; the values of the features $EP_{p1}, \cdots, EP_{p5}$ must be 0. However, when applied to *test* data, these rules may change.

Comparing Table 3 and Table 5, we can see that

- Feature type has been changed. The values of the original features are continuous, but the new features are binary (2-value categorical).
- Distinction between classes is easier. Under the original features' description, the distinction between the two classes is visually difficult. But, in the transformed representation, the distinction is apparent.

3.2 Easy predictions on transformed *test* data

As seen in the above section, under the new features' description, the representation of the transformed training data shows clear distinction between the two classes. The reason is that the

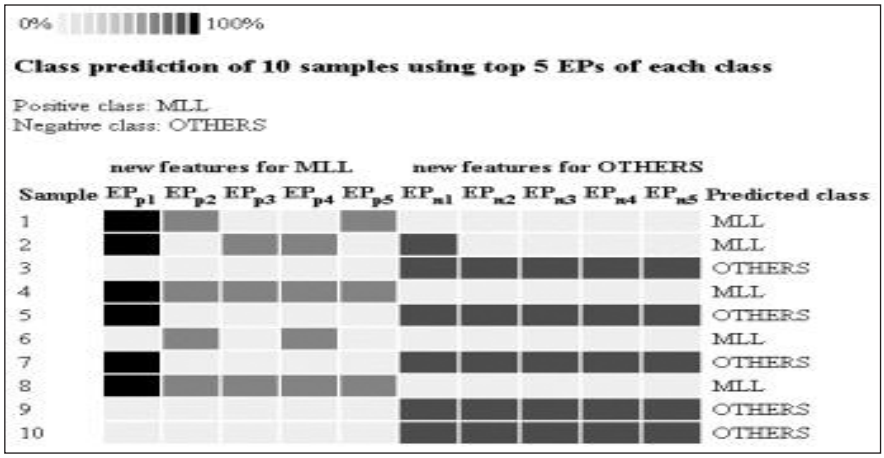

Figure 2: A visual display of 10 test samples under the description of the new features. The frequency information of EPs are incorporated in the figure using different grey scales.

discovery of emerging patterns is based on training data only. Does this distinction happen to test data as well? First, let's look at 10 *test* samples under the description of the same original features (genes). See Table 6. The class of the test sample is unknown, and it is to be predicted. Under this representation, it is hard to make any visual decisions.

Interestingly, under the new features' description, we can see that the 10 test samples can be easily classified. See the last column of Table 7. All the predicted class labels are correct. Note that we did not use distance-based, non-linear kernel function based, or decision tree based classification models. Instead, we took visual decisions, using our eyes only to see how many 1's for MLL and how many 0's for OTHERS. The decisions are intuitive, the reasons are explainable, and the accuracy is high. Let's look at the second MLL test sample of Table 7 to understand the reasons behind our visual decision. Of the 10 new features, this test sample contains 3 MLL features (EP_{p1}, EP_{p3}, and EP_{p4}), and one OTHERS feature EP_{n1}. So, this sample possesses 3 top properties of the MLL class and one property of the OTHERS class. Therefore, we favour MLL as its class label.

For a better visualization, we incorporate frequency information of EPs using different grey scales. Figure 2 is a visual representation of the above same set of test samples for MLL and OTHERS. Each row represents a test sample. The relative shading indicates the "strength" of the presence of an EP in the sample and is calculated from the frequency of occurence of the EPs. The darker the shading, the higher the percentage of occurrence of EP in the training set. From Figure 2, we observe that although EP_{p1} is present strongly in the MLL set, it also appears strongly in 2 of the OTHERS samples as well. So EP_{p1} alone cannot be used to predict MLL. It needs quite a strong presence of EP_{p2} to EP_{p4} for prediction of MLL. The OTHERS class is strongly supported by all its EPs.

4. THE PCL CLASSIFIER

The new representation scheme may sometimes lose its effectiveness at the following tie cases:

- A test sample does not contain any top-five EPs either from the positive class or negative class, but contain lower-ranked EPs.
- A test sample contains top-ranked EPs from both positive and negative classes.

These situations occured in our previous experiments. So, for a sophisticated classification, we have developed the PCL classifier (Li *et al*, 2003) to overcome the confusing situations. The basic

idea of the PCL classifier is to examine the relationship between those top-ranked EPs discovered from training data, and top-ranked EPs that are contained in a test sample. The former EPs are globally ranked; however the latter EPs are sample-specific, local. Local top-ranked EPs may not be globally top-ranked. Based on these local and global EPs, PCL calculates two classification scores. By comparing the two scores, PCL then determines the class of the test sample.

Next we review the PCL classifier. In the subsequent section, we explain classification scores derived by PCL, and use some visualization tools to view the scores such that the prediction results look more transparent.

4.1 The algorithm

Given two classes, D_P and D_N, of data and a testing sample T, the first phase of the PCL classifier is to discover EPs from D_P and D_N. Denote the ranked EPs of D_P as

$$EP_1^{(P)}, EP_2^{(P)}, \cdots, EP_i^{(P)}$$

in descending order of their frequency. Similarly, denote the ranked EPs of D_N as

$$EP_1^{(N)}, EP_2^{(N)}, \cdots, EP_j^{(N)}$$

also in descending order of their frequency. Suppose T contains the following EPs of D_P :

$$EP_{i_1}^{(P)}, EP_{i_2}^{(P)}, \cdots, EP_{i_x}^{(P)}$$

where $i_1 < i_2 < \cdots < i_x \le i$, and the following EPs of D_N :

$$EP_{j_1}^{(N)}, EP_{j_2}^{(N)}, \cdots, EP_{j_y}^{(N)}$$

where $j_1 < j_2 < \cdots < j_y \le j$.

The next step is to calculate two scores for predicting the class label of T. Suppose we use k (k «i and k «j) top-ranked EPs of D_P and D_N. Then we define the score of T in the D_P class as

$$score(T)\_\mathcal{D}_P = \sum_{m=1}^{k} \frac{frequency(EP_{i_m}^{(P)})}{frequency(EP_m^{(P)})}$$

and similarly the score in the D_N class as

$$score(T)\_\mathcal{D}_N = \sum_{m=1}^{k} \frac{frequency(EP_{j_m}^{(N)})}{frequency(EP_m^{(N)})}$$

If $score(T)\_D_P > score(T)\_D_N$, then T is predicted as the class of D_P. Otherwise it is predicted as the class of D_N. Note that

$$0 \le \frac{frequency(EP_{j_m})}{frequency(EP_m)} \le 1$$

Next we demonstrate how the classification scores are computed. Suppose k=5, and the frequencies of the 5 top-ranked EPs of the positive class are sorted as 90% (EP_{p1}), 85% (EP_{p2}), 80% (EP_{p3}), 75% (EP_{p4}), and 70% (EP_{p5}). Assume the test sample T contains EP_{p1} (90%), EP_{p3} (80%), EP_{p5} (70%), EP_{p7} (40%), and EP_{p9} (35%). Then

$$score(T)\_\mathcal{D}_P = \frac{90}{90} + \frac{80}{85} + \frac{70}{80} + \frac{40}{75} + \frac{35}{70}$$

So, $score(T)\_D_P = 3.85$.

Denote

$$\frac{frequency(EP_{j_m}^{(P)})}{frequency(EP_m^{(P)})} = sub\_score_{p_m}$$

and

$$\frac{frequency(EP_{j_m}^{(N)})}{frequency(EP_m^{(N)})} = sub\_score_{n_m}$$

$1 \le m \le k$. Then, each *sub_score* can be viewed as a continuous feature taking values between 0 and 1. So, PCL utilizes another type of feature transformation that is similar to the one discussed in the proceeding sections.

PCL solves the two problems discussed at the beginning of this section by setting k of the scoring formula to be a number like 20 or 25. According to our experience, this heuristic number is good to be around 20.

4.2 High accuracy of the PCL classifier

We report the accuracy of PCL on three biomedical data sets. The first one is about the classification of 7 subtypes of the acute lymphoblastic leukemia (ALL) disease using gene expression profiles (Yeoh *et al*, 2002). The second data set is about the diagnosis of ovarian cancer (Petricoin *et al*, 2002) using proteomic patterns in serum that distinguish ovarian cancer from non-cancer. The proteomic spectra were generated by mass spectroscopy. The data can be found at `http://clinicalproteomics.steem.com`. The third one is used to predict common clinical and pathological phenotypes relevant to the treatment of men diagnosed with prostate cancer (Singh *et al*, 2002). For comparison, we also report the performance of C4.5 (Quinlan, 1993), SVM, and *k*-nearest neighbour (*k*-NN). All these data are also available at our Kent Ridge Bio-medical Data Sets Repository (`http://sdmc.i2r.a-star.edu.sg/GEDatasets/Datasets.html`).

Table 8 shows the error rates of the classifiers on the *test* samples of the first data set when 20 top-ranked genes are selected from the training data.

For the second and third data sets, we have conducted a ten-fold cross validation. PCL, C4.5, SVM, and 3 -NN have made 3, 10, 5, and 5 incorrect predictions respectively over the total 253 ovarian cancer samples; and made 3, 8, 10, and 4 mistakes respectively over the total 102 prostate disease samples.

Overall, we can see that PCL is among the best classifier to provide high accuracy.

5. EXPLANATION OF THE REASONS USED IN PCL

In this section, we present an analysis on the classification scores calculated by PCL. Through the analysis, we can understand deep reasons about the decisions that PCL made. In some cases, PCL may draw our careful attention on those hard test samples due to close classification scores.

The T-ALL subtype is a main subtype of the heterogeneous disease of childhood leukemia (Yeoh *et al*, 2002). Yeoh *et al* (2002) have used only one gene to separate T-ALL against all other subtypes of childhood leukemia. However, some possible rarely-occured human errors on recording data and machine errors by the DNA-chips can happen in the whole process, so it is advisable to use more than one gene to classify T-ALL test samples so that the results are more reliable.

The training data consist of 28 T-ALL samples and 187 samples of other subtypes; the test data consist of 15 T-ALL samples and 97 samples of other subtypes. Running PCL, we obtained classification scores for the 112 test samples, partially shown in Table 9.

Note that PCL obtained these scores by setting its parameter *k* as 10. It means that 10 top-ranked global EPs discovered from the training data and 10 top-ranked local EPs contained in a test sample are considered. Remember that a PCL score is a summation of *k* sub-scores. Each sub-score is between 0 and 1. Therefore, for *k* set as 10, maximally, a classification score is 10; on the other hand, a score can be 0 minimally. So, an ideal *score pair* is 0 (in one class) and *k* (in the other class). In such cases, we can make predictions with very strong confidence.

For example, the second score pair, 10 (T-ALL) and 0 (OTHERS), of Table 9 says that all the top-ranked EPs of the T-ALL class are contained in this test sample, but none of the top-ranked EPs of the OTHERS class is contained. So, we have a very strong confidence to make such a prediction that this test sample is a T-ALL. Of the 112 test samples, 86 predictions have been made based on such ideal score pairs. Note that the average score over the 28 T-ALL training samples is 9.996 (*min* = 9.889, *max* = 10.000), and the average score over the 187 OTHERS training samples is 10.000 (*min* = 9.889, *max* = 10.000).

We have also made one and only one *weak* prediction of the 112 test samples. See the fifth score pair in Table 9. The score for T-ALL is 9.70 and the score for OTHERS is 8.04. They are very close. (The score difference between the two classes for the remaining 111 test samples is at least 8.55.) Although PCL did correctly predict the class label of this test sample, the relative weakness of this decision raises several interesting questions.

Datasets (test data size in each class)	Error Rates for Test Data							
	PCL			C4.5			SVM	3 - N N
	k=20, 25, 30			Single	Bagging	Boosting		
BCR-ABL vs others (6:106)	1:0	1:0	1:0	4:4	6:0	4:4	1:1	1:0
E2A-PBX1 vs others (9:103)	0:0	0:0	0:0	0:0	0:0	0:0	0:0	0:0
HyperL50 vs others (22:90)	2:2	2:2	2:2	4:7	4:2	4:7	0:3	1:4
MLL vs others (6:106)	0:0	0:0	0:0	2:2	1:0	2:2	0:0	0:0
T-ALL vs others (15:97)	0:0	0:0	0:0	0:1	0:1	0:1	0:0	0:0
TEL-AML1 vs others (27:85)	2:0	2:0	2:0	2:2	2:1	2:2	1:1	2:0
minitype vs main (27:85)	11:10	9:12	9:3	10:16	18:0	7:4	9:2	9:1

Table 8: Error rates of the four classifiers on the 112 test samples in the problem of subtype classification of the ALL disease.

Sample	Classification Scores		Real Class	Predicted Class	Decision Confidence
	for T-ALL	for OTHERS			
1	10	0.99	T-ALL	T-ALL	strong
2	10	0	T-ALL	T-ALL	strong
3	10	0	T-ALL	T-ALL	strong
4	10	0	T-ALL	T-ALL	strong
5	**9.70**	**8.04**	**T-ALL**	**T-ALL**	**weak**
6	0	10	OTHERS	OTHERS	strong
7	0	10	OTHERS	OTHERS	strong
8	0	10	OTHERS	OTHERS	strong
9	0	9.99	OTHERS	OTHERS	strong
10	0	10	OTHERS	OTHERS	strong

Table 9: Classification scores calculated by PCL for 10 of the 112 test samples. There is one and only one hard prediction on which we show a weak confidence. However, this weak prediction gives rise to several interesting problems as explained in the text.

1. The two close scores say that this test sample contains not only many outstanding properties of the T-ALL class, but also many outstanding properties of the OTHERS class. This phenomenon appears to be strange because all the other samples (regardless of training or test, T-ALL or OTHERS) contain many significant EPs of only their home class, but few or no EPs from the other class. So the question is, was this patient really suffering from childhood leukemia?
2. If the answer to the first question is affirmative, then we can ask a second question: did this patient relapse after treatment? If the patient did indeed relapse, the medical doctors may need to go back to and check the initial examination records of the patient.
3. If the patient does not relapse, and there are no mistakes in the previous examination records, then is it possible there exists a new subtype under the T-ALL class?

Unfortunately, we are not medical doctors, thus we cannot give satisfactory answers here. All these questions are proposed to help users to understand more about this sample, and to understand some extra use of PCL in classification. It also provides medical doctors many additional insights.

As discussed in Section 4, each *sub_score* can be viewed as a new feature taking values between 0 and 1. Under this description, we can use Figure 3 as a visual display of the sub-scores of the 10 test samples. We can see that the display is easy for us to make decisions and to understand the test data. In Figure 3, the darker the shading, the closer the *sub_score* is to 1. From Figure 3, we can see that in all but test sample 5, the values of the sub-scores are high at 1 or close to 1. Test sample 5 stands out visually as being an outlier from the rest of the samples as almost the entire row is shaded black or dark grey. In such a case, the PCL prediction will be very weak.

The above visualization becomes more useful when the test sample is large (in this case, 112 of them) as it is easier to scan through many rows of data visually than rows of classification scores to identify abnormalities.

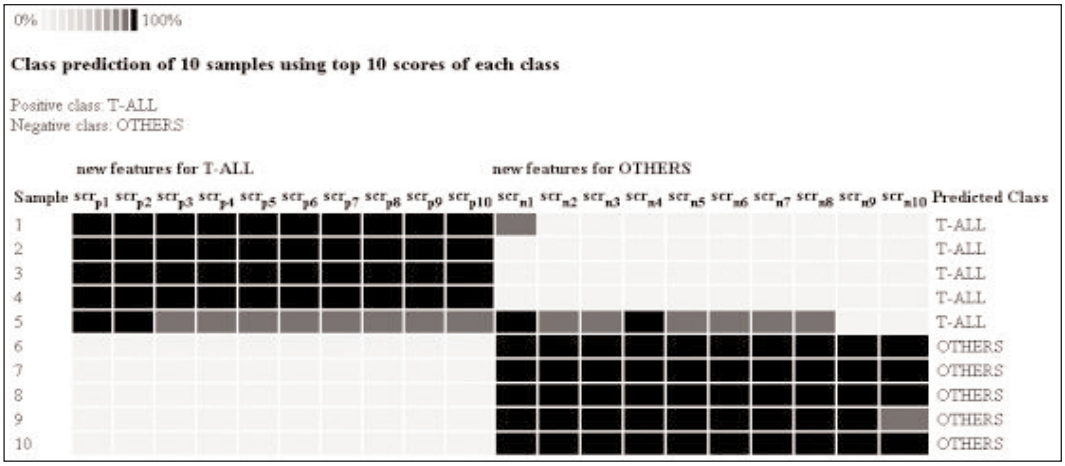

Figure 3: The decomposition of the PCL classification scores under a visual display. This representation can make visual predictions much easier. The symbol *scr* stands for sub-score.

6. CONCLUSION

In this paper, we have discussed the problem of feature transformation. Based on the concept of emerging patterns, we have proposed a new method to combine special groups of original features to form new features. Under the framework of the new features, the original training and test data

are represented in a "tidy" way, having a clear boundary between the two classes. The type of values of the features are changed from continuous to categorical, so that predictions on test data become visually decidable. Meanwhile, the reasons for the predictions are intuitive.

The PCL classifier has shown its high accuracy on the three data sets. Its performance is among the best classifiers. Another advantage of the PCL classifier is that it is a rule-based classifier. Its prediction reasons are easily interpretable. So, more than merely outputting a prediction, it also provides some rule-based insight into the application. The decomposition method on the classification scores can allow us to pay close attention to hard test data. So that we can understand those cases from more than one angle.

Feature transformation, or sometimes called feature generation, and feature selection are two important pre-processes in handling high-dimensional data. Feature selection can narrow our search space, while feature transformation can enhance the expressiveness of the application. Visualization is another important tool for post-data analysis. It can help to explain the data and provide deeper insights. It also helps to quickly zoom in to potentially interesting outliers for further investigations. We will continue our research on these aspects to improve the system.

ACKNOWLEDGMENT

We thank Dr. See-Kiong Ng for providing useful comments on a draft of the paper.

REFERENCES

BURGES, C. (1998): A tutorial on support vector machines for pattern recognition, *Data Mining and Knowledge Discovery* 2: 121–167.

COVER, T.M. and HART, P.E. (1967): Nearest neighbour pattern classification, *IEEE Transactions on Information Theory* 13: 21–27.

DONG, G. and LI, J. (1999): Efficient mining of emerging patterns: Discovering trends and differences, in S. CHAUDHURI and D. MADIGAN, eds, Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM Press, San Diego, CA, 3–52.

FAYYAD, U. and IRANI, K. (1993): Multi-interval discretization of continuous-valued attributes for classification learning, in R. BAJCSY, ed., Proceedings of the Thirteenth International Joint Conference on Artificial Intelligence, Morgan Kaufmann, 1022–1029.

LI, J., LIU, H., DOWNING, J.R., YEOH, A.E-J. and WONG, L. (2003): Simple rules underlying gene expression profiles of more than six subtypes of acute lymphoblastic leukemia (ALL) patients, *Bioinformatics* 19: 71–78.

LI, J., RAMAMOHANARAO, K. and DONG, G. (2000): The space of jumping emerging patterns and its incremental maintenance algorithms, in Proceedings of the Seventeenth International Conference on Machine Learning, Stanford, CA, USA, Morgan Kaufman, San Francisco, 551–558.

LI, J and WONG, L. (2002a): Geography of differences between two classes of data, in Proceedings of the Sixth European Conference on Principles of Data Mining and Knowledge Discovery, PKDD 2002, Springer-Verlag, Helsinki, Finland, 325–337.

LI, J. and WONG, L. (2002b): Identifying good diagnostic gene groups from gene expression profiles using the concept of emerging patterns, *Bioinformatics* 18: 725–734.

PETRICOIN, E.F., ARDEKANI, A.M., HITT, B.A., LEVINE, P.J., FUSARO, V.A., STEINBERG, S.M., MILLS, G.B., SIMONE, C., FISHMAN, D.A., KOHN, E.C. and LIOTTA, L.A. (2002): Use of proteomic patterns in serum to identify ovarian cancer, *Lancet* 359: 572–577.

QUINLAN, J.R. (1993): *C4.5: Programs for Machine Learning*, Morgan Kaufmann, San Mateo, CA.

SINGH, D., FEBBOL, P.G., ROSS, K., JACKSON, D.G., MANOLA, J., LADD, C., TAMAYO, P., RENSHAW, A.A., D'AMICO, A.V., RICHIE, J.P., LANDER, E.S., LODA, M., KANTOFF, P.W., GOLUB, T.R. and SELLERS, W.R. (2002): Gene expression correlates of clinical prostate cancer behaviour, *Cancer Cell* 1: 203–209.

YEOH, E.J., ROSS, M.E., SHURTLEFF, S.A., WILLIAMS, W.K., PATEL, D., MAHFOUZ, R., BEHM, F.G., RAIMONDI, S.C., RELLING, M.V., PATEL, A., CHENG, C., CAMPANA, D., WILKINS, D., ZHOU, X., LI, J., LIU, H., PUI, C-H., EVANS, W.E., NAEVE, C., WONG, L. and DOWNING, J.R. (2002): Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling, *Cancer Cell* 1: 133–143.

BIOGRAPHICAL NOTES

Jinyan Li is senior scientist at the Institute for Infocomm Research, Singapore. He got his PhD in computer science from the University of Melbourne, Australia in 2001. His recent research interests include machine learning, data mining, bioinformatics (gene expression and proteomic profiling data analysis, clinical data analysis, mutation resistance studies, and protein-protein interactions), and decision systems.

Jinyan Li

Hwee-Leng Ong is a principal research engineer at the Institute for Infocomm Research, Singapore. She received her BSc (Computer Science) in 1989 at the University of Southern California. Her research interest includes information visualization, text mining and data mining as well as industrial projects leading to the deployment or productization of the research. These include a user-configurable clustering tool for competitive intelligence, a data mining visualization tool as well as an internet based expert system. She has also served on the organizing committees of PAKDD97 and PRICAI98. She is a member of ACM, SIGKDD and the Singapore Computer Society (SCS).

Hwee-Leng Ong