Geography of Differences between Two Classes of Data

Jinyan Li[1] and Limsoon Wong[1]

Laboratories for Information Technology, 21 Heng Mui Keng Terrace, Singapore 119613.
{jinyan, limsoon}@lit.org.sg

Abstract. Easily comprehensible ways of capturing main differences between two classes of data are investigated in this paper. In addition to examining individual differences, we also consider their neighbourhood. The new concepts are applied to three gene expression datasets to discover diagnostic gene groups. Based on the idea of prediction by collective likelihoods (PCL), a new method is proposed to classify testing samples. Its performance is competitive to several state-of-the-art algorithms.

1 Introduction

An important problem in considering two classes of data is to discover significant differences between the two classes. This type of knowledge is useful in biomedicine. For example, in gene expression experiments [1, 6], doctors and biologists wish to know genes or gene groups whose expression levels change sharply between normal cells and disease cells. Then, these genes or their protein products can be used as diagnostic indicators or drug targets of that specific disease.

Based on the concept of emerging patterns [3], we define a *difference* as a set of conditions that most data of a class satisfy but none of the other class satisfy. We investigate the geography—properties of neighbourhoods—of these differences. The differences include those corresponding to boundary rules for separating the two classes, those at the same level of significance in one class, and those at lower part of the boundaries. After examining these neighbourhoods, we can identify differences that are more interesting. We first discuss our ideas in a general sense. Then we apply the methods to three gene expression datasets [1, 6] to discover interesting gene groups. We also use the discovered patterns to do classification and prediction.

Suppose we are given two sets of relational data where a fixed number of *features* (also called *attributes*) exist. Every feature has a range of numeric real values or a set of categorical values. A *condition* (also called *item*) is defined as a pair of a feature and its value. An example of a condition (an item) is "the expression of $gene\_x$ is less than 1000". We denote this condition by $gene\_x@(-\infty, 1000)$, where the feature is $gene\_x$ and its value is $(-\infty, 1000)$. An *instance* (or a sample) is defined as a set of conditions (items) with a cardinality equal to the number of features in the relational data.

A *pattern* is a set of conditions. A pattern is said to *occur* in an instance if the instance contains it. For two classes of instances, a pattern can have a very high occurrence (equivalently, frequency) in one class, but can change to a low or even zero occurrence in the other class. Those patterns with a significant occurrence change are called emerging patterns (EPs) [3]. Here, our differences are those described by EPs.

This paper is organized as follows: Firstly, we present a formal description of the problems, including the definition of *boundary* EPs, *plateau spaces*, and *shadow patterns*, and present a related work. Then we describe convex spaces and prove that all plateau spaces satisfy convexity. This property is useful in concisely representing large pattern spaces. We also try to categorize boundary EPs using the frequency of their subsets. Then we present our main results, patterns discovered from biological data, and explain them in both biological and computational ways. To show the potential of our patterns in classification, we propose a new method that sums the collective power of individual patterns. Our accuracy is better than other methods. Then we briefly report our recent progress on a very big gene expression dataset which is about the subtype classification and relapse study of Acute Lymphoblastic Leukemia.

2 Problems and Related Work

Three types of patterns—*boundary* EPs, *plateau* EPs, and *shadow patterns*—are investigated in this work. Let us begin with a definition of emerging patterns.

Definition 1. *Given two classes of data, an* emerging pattern *is a pattern whose frequency in one class is non-zero but in the other class is zero.*

Usually, the class in which an EP has a non-zero frequency is called the EP's *home* class or its own class. The other class in which the EP has the zero frequency is called the EP's *counterpart* class.

2.1 Boundary EPs

Many EPs may have very low frequency (e.g. 1 or 2) in their home class. So boundary EPs are proposed to capture big differences between the two classes:

Definition 2. *A* boundary *EP is an EP whose proper subsets are not EPs.*

How do boundary EPs capture big differences? If a pattern contains less number of items (conditions), then the frequency (probability) that it occurs in a class becomes larger. Removing any one item from a boundary EP thus increases its home class frequency. However, by definition of boundary EPs, the frequency of any of its subsets in the counterpart class must be non-zero. Therefore, boundary EPs are maximally frequent in their home class. They separate EPs from non-EPs. They also distinguish EPs with high occurrence from EPs with low occurrence.

Efficient discovery of boundary EPs has been solved in our previous work [12]. Our new contribution in this work is the ranking of boundary EPs. The number of boundary EPs is sometimes large. The top-ranked patterns can help users understand applications better and easier. We also propose a new algorithm to make use of the frequency of the top-ranked patterns for classification.

2.2 Plateau EPs and Plateau Spaces

Next we discuss a new type of emerging patterns. If one more condition (item) is added to a boundary EP, generating a superset of the EP, the new EP may still have the same frequency as the boundary EP's. We call those EPs having this property plateau EPs:

Definition 3. *Given a boundary EP, all its supersets having the same frequency are called its* plateau *EPs.*

Note that boundary EPs themselves are trivially their plateau EPs. Next we define a new space, looking at all plateau EPs as a whole.

Definition 4. *All plateau EPs of all boundary EPs with the same frequency are called a* plateau space *(or simply, a* P-space*).*

So, all EPs in a P-space are at the same significance level in terms of their occurrence in both their home class and counterpart class. Suppose the home frequency is n, then the P-space is specially denoted P_n*-space*.

We will prove that all P-spaces have a nice property called *convexity*. This means a P-space can be succinctly represented by its most *general* and most *specific* elements. [1] We study how P-spaces contribute to the high accuracy of our classification system.

2.3 Shadow Patterns

All EPs defined above have the same *infinite* frequency growth-rate from their counterpart class to their home class. However, all proper subsets of a boundary EP have a *finite* frequency growth-rate as they occur in both the classes. It is interesting to see how these subsets change their frequency between the two classes by studying the growth rates. Next we define shadow patterns, which are special subsets of a boundary EP.

Definition 5. *All immediate subsets of a boundary EP are called* shadow patterns.

Shadow patterns can be used to measure the interestingness of boundary EPs. Given a boundary EP X, if the growth-rates of its shadow patterns approach $+\infty$, then the existence of this boundary EP is reasonable. This is because the possibility of X being a boundary EP is large. Otherwise if the growth-rates of the shadow patterns are on average around small numbers like 1 or 2, then the pattern X is *adversely interesting*. This is because the possibility of X being a boundary EP is small; the existence of this boundary EP is "unexpected". This conflict may reveal some new insights into the correlation of the features.

2.4 Related Work on EPs

The general discussion of EP spaces has been thoroughly studied in our earlier work [12]. It has been proven that every EP space is a convex space. The efficient discovery of boundary EPs was a problem and it was solved by using border-based algorithms [3,

[1] Given a collection $\mathcal{C}$ of patterns and $A \in \mathcal{C}$, A is most general if there is no proper subset of A in $\mathcal{C}$. Similarly, A is most specific if there is no proper superset of A in $\mathcal{C}$.

12]. Based on experience, the number of boundary EPs is usually large— from 100s to 1000s depending on datasets. So, the ranking and visualization of these patterns is an important issue. We propose some ideas here to sort and list boundary EPs.

The original idea of the concept of emerging patterns is proposed in [3]. General definition of EPs, its extension to spatial data and to time series data, and the mining of general EPs can be also found there [3]. This paper discusses two new types of patterns: plateau patterns and shadow patterns. They are closely related to boundary EPs. We study these three types of patterns together here.

The usefulness of EPs in classification has been previously investigated [4, 11]. We propose in this paper a new idea that only top-ranked boundary EPs are used in classification instead of using all boundary EPs. This new idea leads to a simple system without any loss of accuracy and can avoid the effect of possible noisy patterns.

3 The Convexity of P-spaces

Convexity is an important property of a certain type of large collections. It can be exploited to concisely represent those collections of large size. Next we give a definition of convex space. Then we prove that our P-spaces satisfy convexity.

Definition 6. *A collection $\mathcal{C}$ of patterns is a* convex space *if, for any patterns X, Y, and Z, the conditions $X \subseteq Y \subseteq Z$ and $X, Z \in \mathcal{C}$ imply that $Y \in \mathcal{C}$.*

If a collection is a convex space, it is said to hold *convexity*. More discussion about convexity can be found in [7].

Example 1. The patterns $\{a\}$, $\{a, b\}$, $\{a, c\}$, $\{a, d\}$, $\{a, b, c\}$, and $\{a, b, d\}$ form a convex space. The set $\mathcal{L}$ consisting of the most general elements in this space is $\{\{a\}\}$. The set $\mathcal{R}$ consisting of the most specific elements in this space is $\{\{a, b, c\}, \{a, b, d\}\}$. All the other elements can be considered to be "between" $\mathcal{L}$ and $\mathcal{R}$.

Theorem 1. *Given a set $\mathcal{D}_P$ of positive instances and a set $\mathcal{D}_N$ of negative instances, every P$_n$-space $(n \geq 1)$ is a convex space.*

Proof. By definition, a P$_n$-space is the set of all plateau EPs of all boundary EPs with the same frequency of n in the same home class. Without loss of generality, suppose two patterns X and Z satisfy (i) $X \subseteq Z$; (ii) X and Z are plateau EPs having the occurrence of n in $\mathcal{D}_P$. Then, for any pattern Y satisfy $X \subseteq Y \subseteq Z$, it is a plateau EP with the same n occurrence in $\mathcal{D}_P$. This is because

1. X does not occur in $\mathcal{D}_N$. So, Y, a superset of X, does not occur in $\mathcal{D}_N$ either.
2. The pattern Z has n occurrences in $\mathcal{D}_P$. So, Y, a subset of Z, also has a non-zero frequency in $\mathcal{D}_P$.
3. The frequency of Y in $\mathcal{D}_P$ must be less than or equal to the frequency of X, but must be larger than or equal to the frequency of Z. As the frequency of both X and Z is n, the frequency of Y in $\mathcal{D}_P$ is also n.
4. X is a superset of a boundary EP, thus Y is a superset of some boundary EP as $X \subseteq Y$.

By the first two points, we can infer that Y is an EP of $\mathcal{D}_P$. From the third point, we know that Y's occurrence in $\mathcal{D}_P$ is n. Therefore, with the forth point above, Y is a plateau EP. Then we have proven that every P_n-space is a convex space.

A plateau space can be bounded by two sets similar to the sets $\mathcal{L}$ and $\mathcal{R}$ as shown in example 1. The set $\mathcal{L}$ consists of the boundary EPs. These EPs are the most general elements of the P-space. Usually, features contained in the patterns in $\mathcal{R}$ are more numerous than the patterns in $\mathcal{L}$. This indicates that some feature groups can be expanded while keeping their significance.

The structure of an EP space can be understood in a way by decomposing the space into a series of P-spaces and a non P-space. This series of P-spaces can be sorted according to their frequency. Interestingly, one of them with the highest frequency is a version space [14, 8] if the EPs have the full 100% frequency in their home class.

4 Our Discovered Patterns from Gene Expression Datasets

We next apply our methods to two public datasets. One contains gene expression levels of normal cells and cancer cells. The other contains gene expression levels of two main subtypes of a disease. We report our discovered patterns, including boundary EPs, P-spaces, and shadow patterns. We also explain these patterns in a biological sense.

4.1 Data Description

The process of transcribing a gene's DNA sequence into RNA is called *gene expression*. After translation, RNA becomes proteins consisting of amino-acid sequences. A gene's expression level is the rough number of copies of that gene's RNA produced in a cell.

Gene expression data, obtained by highly parallel experiments using technologies like oligonucleotide 'chips' [13], record expression levels of genes under specific experimental conditions. By conducting gene expression experiments, one hopes to find possible trends or regularities of every single gene under a series of conditions, or to identify genes whose expressions are good diagnostic indicators for a disease.

A leukemia dataset [6] and a colon tumor dataset [1] are used in this paper. The former contains a training set of 27 samples of acute lymphoblastic leukemia (ALL) and 11 samples of acute myeloblastic leukemia (AML), and a blind testing set of 20 ALL and 14 AML samples. (ALL and AML are two main subtypes of the leukemia disease.) The high-density oligonucleotide microarrays used 7129 probes of 6817 human gene. All these data are public available at `http://www.genome.wi.mit.edu/MPR`. The second dataset consists of 22 normal and 40 colon cancer tissues. The expression level of 2000 genes of these samples are recorded. The data is available at `http://microarray.princeton.edu/oncology/affydata/index.html`. We use Table 1 to summarize the data. A common characteristic of gene expression data is that the number of samples is not large and the number of features is high in comparison with commercial market data.

Table 1. Two publicly accessible gene expression datasets.

Dataset	Gene number	Training size	Classes
Leukemia	7129	27, 11	ALL, AML
Colon	2000	22, 40	Normal, Cancer

4.2 Gene Selection and Discretization

A major challenge in analysing gene expression data is the overwhelming number of features. How to extract informative genes and how to avoid noisy data effects are important issues. We use an entropy-based method [5, 9] and the CFS (Correlation-based Feature Selection) algorithm [16] to perform feature selection and discretization.

The entropy-based discretization method ignores those features which contain a random distribution of values with different class labels. It finds those features which have big intervals containing almost the same class of points. The CFS method is a post-process of the discretization. Rather than scoring (and ranking) individual features, the method scores (and ranks) the worth of subsets of the discretized features [16].

4.3 Patterns Derived from the Leukemia Data

The CFS method selects only one gene, Zyxin, from the total of 7129 features. The discretization method partitions this feature into two intervals using the cut point at 994. Then, we discovered two boundary EPs,

$$\{gene\_zyxin@(-\infty, 994)\} \text{ and } \{gene\_zyxin@[994, +\infty)\},$$

having a 100% occurrence in their home class.

Biologically, these two EPs say that if the expression of Zyxin in a cell is less than 994, then this cell is an ALL sample. Otherwise this cell is an AML sample. This rule regulates all 38 training samples without any exception. If this rule is applied to the 34 blind testing samples, we obtained only three mis-classifications. This result is better than the accuracy of the system reported in [6].

Biological and technical noise sometimes happen in many stages such as in the production of DNA arrays, the preparation of samples, the extraction of expression levels, and may be from the impurity or mis-classification of tissues. To overcome these possible machine and human minor errors, we suggest to use more than one gene to strengthen our system as shown later.

We found four genes whose entropy values are significantly less than all the other 7127 features when partitioned by the discretization method. We used these four genes for our pattern discovery whose name, cut points, and item indexes are listed in Table 2.

We discovered a total of 6 boundary EPs, 3 each in the ALL and AML classes. Table 3 presents the boundary EPs together with their occurrence and the percentage of the occurrence in the whole class. The reference numbers contained in the patterns can be interpreted using the interval index in Table 2.

Biologically, the EP $\{5, 7\}$ as an example says that if the expression of CST3 is less than 1419.5 and the expression of Tropomysin is less than 83.5 then this sample is ALL

Table 2. Four most discriminatory genes of the 7129 features. Each feature is partitioned into two intervals using the cut points in column 2. The item index is convenient for writing EPs.

Features	Cut Point	Item Index
Zyxin	994	1, 2
FAH	1346	3, 4
CST3	1419.5	5, 6
Tropomyosin	83.5	7, 8

Table 3. Three boundary EPs in the ALL class and three boundary EPs in the AML class.

Boundary EPs	Occurrence in ALL (%)	Occurrence in AML (%)
$\{5, 7\}$	27 (100%)	0
$\{1\}$	27 (100%)	0
$\{3\}$	26 (96.3%)	0
$\{2\}$	0	11 (100%)
$\{8\}$	0	10 (90.9%)
$\{6\}$	0	10 (90.9%)

with 100% accuracy. So, all those genes involved in our boundary EPs are very good diagnostic indicators for classifying ALL and AML.

We discovered a P-space based on two boundary EPs of $\{5, 7\}$ and $\{1\}$. This P_{27}-space consists of five plateau EPs: $\{1\}$, $\{1, 7\}$, $\{1, 5\}$, $\{5, 7\}$, and $\{1, 5, 7\}$. The most specific plateau EP is $\{1, 5, 7\}$ and it still has a full occurrence of 27 in the ALL class.

4.4 Patterns Derived from the Colon Tumor Data

This dataset is a bit more complex than the ALL/AML data. The CFS method selected 23 features from the 2000 as most important. All of the 23 features were partitioned into two intervals.

We discovered 371 boundary EPs in the normal cells class, and 131 boundary EPs in the cancer cells class. The total 502 patterns were ranked according to the these criteria:

1. Given two EPs X_i and X_j, if the frequency of X_i is larger than X_j, then X_i is prior to X_j in the list.
2. When the frequency of X_i is equal to X_j, if the cardinality of X_i is larger than X_j, then X_i is prior to X_j in the list.
3. If their frequency and cardinality are both identical, then X_i is prior to X_j when X_i is first produced.

Some top ranked boundary EPs are reported in Table 4.

Unlike the ALL/AML data, in the colon tumor dataset there does not exist single genes acting as arbitrator to separate normal and cancer cells clearly. Instead, gene groups are contrasting the two classes. Note that these boundary EPs, especially those having many conditions, are not obvious but novel to biologists and medical doctors. They may reveal some new protein interactions and may be used to find new pathways.

Table 4. Here only top 5 ranked boundary EPs in the normal class and in the cancerous class are listed. The meaning of the reference numbers contained in the patterns are not presented due to page limitation.

Boundary EPs	Occurrence Normal (%)	Occurrence Cancer (%)
$\{2, 6, 7, 11, 21, 23, 31\}$	18 (81.8%)	0
$\{2, 6, 7, 21, 23, 25, 31\}$	18 (81.8%)	0
$\{2, 6, 7, 9, 15, 21, 31\}$	18 (81.8%)	0
$\{2, 6, 7, 9, 15, 23, 31\}$	18 (81.8%)	0
$\{2, 6, 7, 9, 21, 23, 31\}$	18 (81.8%)	0
$\{14, 34, 38\}$	0	30 (75.0%)
$\{18, 34, 38\}$	0	26 (65.0%)
$\{18, 32, 38, 40\}$	0	25 (62.5%)
$\{18, 32, 44\}$	0	25 (62.5%)
$\{20, 34\}$	0	25 (62.5%)

Table 5. Most general and most specific elements in a P_{18}-space in the normal class of the colon data.

Most general and specific EPs	Occurrence in Normal
$\{2, 6, 7, 11, 21, 23, 31\}$	18
$\{2, 6, 7, 21, 23, 25, 31\}$	18
$\{2, 6, 7, 9, 15, 21, 31\}$	18
$\{2, 6, 7, 9, 15, 23, 31\}$	18
$\{2, 6, 7, 9, 21, 23, 31\}$	18
$\{2, 6, 9, 21, 23, 25, 31\}$	18
$\{2, 6, 7, 11, 15, 31\}$	18
$\{2, 6, 11, 15, 25, 31\}$	18
$\{2, 6, 15, 23, 25, 31\}$	18
$\{2, 6, 15, 21, 25, 31\}$	18
$\{2, 6, 7, 9, 11, 15, 21, 23, 25, 31\}$	18

There are a total of ten boundary EPs having the same highest occurrence of 18 in the normal cells class. Based on these boundary EPs, we found a P_{18}-space in which the only most specific element is $Z = \{2, 6, 7, 9, 11, 15, 21, 23, 25, 31\}$. By convexity, any subsets of Z but superset of anyone of the ten boundary EPs have the occurrence of 18 in the normal class. Observe that there are approximately one hundred EPs in this P-space. While by convexity, we can concisely represent this space using only 11 EPs which are shown in Table 5.

From this P-space, it can be seen that significant gene groups (boundary EPs) can be expanded by adding some other genes without loss of significance, namely still keeping high occurrence in one class but absence in the other class. This may be useful in identifying a maximum length of a pathway.

We found a P_{30}-space in the cancerous class. The only most general EP in this space is $\{14, 34, 38\}$ and the only most specific EP is $\{14, 30, 34, 36, 38, 40, 41, 44, 45\}$. So a boundary EP can be extended by six more genes without a reduction in occurrence.

Table 6. A boundary EPs and its three shadow patterns.

Patterns	Occurrence in Normal	Occurrence in Cancer
$\{14, 34, 38\}$	0	30
$\{14, 34\}$	1	30
$\{14, 38\}$	7	38
$\{34, 38\}$	5	31

It is easy to find shadow patterns. Below, we report a boundary EP and its shadow patterns (see Table 6). These shadow patterns can also be used to illustrate the point that proper subsets of a boundary EP must occur in two classes at non-zero frequency.

5 Usefulness of EPs in Classification

In the previous section, we have found many simple EPs and rules which can well regulate gene expression data. Next we propose a new method, called PCL, to test the reliability and classification potential of the patterns by applying them to the 34 blind testing sample of the leukemia dataset [6] and by conducting a Leave-One-Out cross-validation (LOOCV) on the colon dataset.

5.1 Prediction by Collective Likelihood (PCL)

From the leukemia training data, we first discovered two boundary EPs which form a simple rule. So, there was no ambiguity in using the rule. However, a large number of EPs were found in the colon dataset. A testing sample may contain not only EPs from its own class, but it may also contain EPs from its counterpart class. This makes the prediction a bit more complicated. Naturally, a testing sample should contain many top-ranked EPs from its own class and contain a few low-ranked, preferably no, EPs from its opposite class. However, according to our observations, a testing sample can sometimes, though rarely, contain 1 to 20 top-ranked EPs from its counterpart class. To make reliable predictions, it is reasonable to use multiple highly frequent EPs of the home class to avoid the confusing signals from counterpart EPs.

Our method is described as follows: Given two training datasets $\mathcal{D}_P$ and $\mathcal{D}_N$ and a testing sample T, the first phase of our prediction method is to discover boundary EPs from $\mathcal{D}_P$ and $\mathcal{D}_N$. Denote the ranked EPs of $\mathcal{D}_P$ as,

$$TopEP\_P\_1, TopEP\_P\_2, \cdots, TopEP\_P\_i,$$

in descending order of frequency. Similarly, denote the ranked boundary EPs of $\mathcal{D}_N$ as

$$TopEP\_N\_1, TopEP\_N\_2, \cdots, TopEP\_N\_j$$

also in descending order of frequency. Suppose T contains the following EPs of $\mathcal{D}_P$:

$$TopEP\_P\_i_1, TopEP\_P\_i_2, \cdots, TopEP\_P\_i_x,$$

where $i_1 < i_2 < \cdots < i_x \leq i$, and the following EPs of $\mathcal{D}_N$:

$$TopEP\_N\_j_1, TopEP\_N\_j_2, \cdots, TopEP\_N\_j_y,$$

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

The next step is to calculate two scores for predicting the class label of T. Suppose we use k ($k \ll i$ and $k \ll j$) top-ranked EPs of $\mathcal{D}_P$ and $\mathcal{D}_N$. Then we define the score of T in the $\mathcal{D}_P$ class as

$$score(T)\_\mathcal{D}_P = \sum_{m=1}^{k} \frac{frequency(TopEP\_P\_i_m)}{frequency(TopEP\_P\_m)},$$

and similarly the score in the $\mathcal{D}_N$ class as

$$score(T)\_\mathcal{D}_N = \sum_{m=1}^{k} \frac{frequency(TopEP\_N\_j_m)}{frequency(TopEP\_N\_m)}.$$

If $score(T)\_\mathcal{D}_P > score(T)\_\mathcal{D}_N$, then T is predicted as the class of $\mathcal{D}_P$. Otherwise predicted as the class of $\mathcal{D}_N$. We use the size of $\mathcal{D}_P$ and $\mathcal{D}_N$ to break tie.

The spirit of our proposal is to measure how far the top k EPs contained in T are away from the top k EPs of a class. Assume $k = 1$, then $score(T)\_\mathcal{D}_P$ indicates whether the number one EP contained in T is far from the most frequent EP of $\mathcal{D}_P$. If the score is the maximum value 1, then the "distance" is very close, namely the most common property of $\mathcal{D}_P$ is also present in this testing sample. With smaller scores, the distance becomes further. Thus the likelihood of T belonging to the class of $\mathcal{D}_P$ becomes weaker. Using more than one top-ranked EPs, we utilize a "collective" likelihood for more reliable predictions. We name this method PCL (prediction by collective likelihood).

5.2 Classification Results

Recall that we also have selected four genes in the leukemia data as the most important. Using PCL, we obtained a testing error rate of two mis-classifications. This result is one error less than the result obtained by using the sole Zyxin gene.

For the colon dataset, using our PCL, we can get a better LOOCV error rate than other classification methods such as C4.5 [15], Naive Bayes (NB) [10], k-NN, and support vector machine (SVM) [2]. We used the default settings of the Weka package [16] and exactly the gene selection preprocessing steps as ours to get the results. The result is summarized in Table 7.

5.3 Making Use of P-spaces for Classification: A Variant of PCL

Can the most specific elements of P-spaces be useful in classification? In PCL, we tried to replace the ranked boundary EPs with the most specific elements of all P-spaces in the colon dataset. The remaining process of PCL are not changed. By LOOCV, we obtained an error rate of only six mis-classifications. This reduction is significant.

Table 7. By LOOCV on the colon dataset, our PCL's error rate comparison with other methods.

Methods	Error Rates
C4.5	20
NB	13
k-NN	28
SVM	24
Our PCL	13, 12, 10, 10, 10, 10
$(k = 5, 6, 7, 8, 9, 10)$	

The reason for this good result is that the neighbourhood of the most specific elements of a P-space are all EPs in most cases, but there are many patterns in the neighbourhood of boundary EPs that are *not* EPs. Secondly, the conditions contained in the most specific elements of a P-space are usually much more than the boundary EPs. So, with more number of conditions, the chance for a testing sample to contain opposite EPs becomes smaller. Hence, the probability of being correctly classified becomes higher.

6 Recent Progress

In a collaboration with St. Jude Children's Research Hospital, our algorithm has been applied to a big gene expression dataset [17]. This dataset consists of the expression profile of 327 patients who suffered from Acute Lymphoblastic Leukemia (ALL). Each instance is represented by 12,558 features. The purpose is to establish a classification model to predict whether a new patient suffers from one of the six main subtypes of ALL. By our PCL, we achieved a testing error rate that is 71% better than C4.5, 50% better than Naive Bayes, 43% better than k-NN, and 33% better than SVM.

More than mere a prediction, importantly, our algorithm provides simple rules and patterns. These knowledge can greatly help medical doctors and biologists deeply understand why an instance is predicted as positive or negative.

7 Conclusion

We studied how to describe main differences between two classes of data using emerging patterns. We proposed methods to rank boundary EPs. Using boundary EPs, we defined two new types of patterns, plateau EPs and shadow patterns, and proved that all P-spaces satisfied convexity. Based on the idea of prediction by collective likelihood, we proposed a new classification method called PCL.

All these ideas and methods have been applied to three gene expression data. The discovered patterns are interesting, and may be useful in identifying new pathways and interactions between proteins. The PCL methods performed better than other classification models on the datasets used in this paper.

In future, we plan to define central points of a P-space and use the central patterns for classification. Also, we like to study shadow patterns and their relation with boundary EPs more deeply than in this paper.

Acknowledgments

We thank Huiqing Liu for providing the classification results of C4.5, NB, k-NN, and SVM. We also thank the reviewers for their useful comments.

References

1. Alon, U. and et al. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proceedings of National Academy of Sciences of the United States of American*, 96:6745–675, 1999.
2. Burges, C.J. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2:121–167, 1998.
3. Guozhu Dong and Jinyan Li. Efficient mining of emerging patterns: Discovering trends and differences. In *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 43–52, San Diego, CA, 1999. ACM Press.
4. Guozhu Dong, Xiuzhen Zhang, Limsoon Wong, and Jinyan Li. CAEP: Classification by aggregating emerging patterns. In *Proceedings of the Second International Conference on Discovery Science, Tokyo, Japan*, pages 30–42. Springer-Verlag, December 1999.
5. Fayyad, U.M. and Irani, K.B. Multi-interval discretization of continuous-valued attributes for classification learning. In *Proceedings of the 13th International Joint Conference on Artificial Intelligence*, pages 1022–1029. Morgan Kaufmann, 1993.
6. Golub, T.R. and *et al*. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, 286:531–537, October 1999.
7. Carl A. Gunter, Teow-Hin Ngair, and Devika Subramanian. The common order-theoretic structure of version spaces and ATMS's. *Artificial Intelligence*, 95:357–407, 1997.
8. Hirsh, H. Generalizing version spaces. *Machine Learning*, 17:5–46, 1994.
9. Kohavi, R. and *et al*. MLC++: A machine learning library in C++. In *Tools with artificial intelligence*, pages 740 – 743, 1994.
10. Langley, P., Iba, W. and Thompson, K. An analysis of Bayesian classifier. In *Proceedings of the Tenth National Conference on Artificial Intelligence*, pages 223 – 228. AAAI Press, 1992.
11. Jinyan Li, Guozhu Dong, and Kotagiri Ramamohanarao. Making use of the most expressive jumping emerging patterns for classification. *Knowledge and Information Systems: An International Journal*, 3:131–145, 2001.
12. Jinyan Li, Kotagiri Ramamohanarao, and Guozhu Dong. The space of jumping emerging patterns and its incremental maintenance algorithms. In *Proceedings of the Seventeenth International Conference on Machine Learning, Stanford, CA, USA*, pages 551–558, San Francisco, June 2000. Morgan Kaufmann.
13. Lockhart, T.J. and *et al*. Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nature Biotechnology*, 14:1675–1680, 1996.
14. Mitchell, T.M. Generalization as search. *Artificial Intelligence*, 18:203–226, 1982.
15. Quinlan, J.R. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, San Mateo, CA, 1993.
16. Witten, H. and Frank, E. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementation*. Morgan Kaufmann, San Mateo, CA, 2000.
17. Eng-Juh Yeoh and *et. al.* Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling. *Cancer Cell*, 1:133–143, 2002.