

Using Emerging Pattern Based Projected Clustering and Gene Expression Data for Cancer Detection

Larry T.H. Yu, Fu-lai Chung, Stephen C.F. Chan and Simon M.C. Yuen

Department of Computing
The Hong Kong Polytechnic University
Hung Hom, Kowloon, Hong Kong

{csthyu, cskchung, csschan, csmcyuen}@comp.polyu.edu.hk

Abstract

Using gene expression data for cancer detection is one of the famous research topics in bioinformatics. Theoretically, gene expression data is capable to detect all types of early cancer development in molecular level. Traditional clustering and pattern mining algorithm are either inadequate to handle high dimensional gene expression data effectively or the results obtained are not easy to understand. We proposed emerging pattern based projected clustering (EPPC) approaches to cope with the cancer detection problem. Previous result shows that easy understandable clusters are obtained. In this paper, the dimension projection process of EPPC is further studied and experimental results showed that the resulting clusters obtained by EPPC give comparable accuracy in classification when compared with ORCLUS.

Keywords: Data mining, emerging patterns, projected clustering, bioinformatics, gene expression data and cancer classification.

1 Introduction

Cancer detection is one of the important research topics in medical science. It is because cancer is one of the top killer diseases, our knowledge in cancer development is limited and the number of new patients is still increasing in recently years. In general, the survival rate of patients is much higher if cancers are detected and treatments are started at its early stage. Therefore, cancer detection techniques are very important for everyone and it is especially useful for the doctors and biologists in diagnosis and new treatment plans discovery.

Cancer detection is not an easy task since early cancer may not have any symptoms. Most of the patients only visit their doctors when they noticed changes occur in their bodies, such as lump in breast. When such changes are noticeable, the development of cancers is no longer at the early stage. Thus, screening methods that designed to check for cancer in people with no symptoms are

developed. Pap test, mammogram, PSA test or fecal occult blood test (NCI 2003b) are currently used to indicate the possible existence of cancer. However, our knowledge about cancer is still limited, those screening methods are not sufficient to detect all types of cancers and many cancers cannot be detected accurately in their early stage (NCI 2003b).

In bioinformatics age, gene expression data can be used for the cancer detection. Chemicals, radiation, viruses and heredity all contributed to the development of cancer by changing the genetic information in our cells. Genes are the inherited instructions that are regions within our DNA molecules will turn on, also called gene express, and produce protein. In these protein synthesis processes, different genes expressed differently and scientists used microarray technology to record the values of gene expression in large scale. They found that context of gene expression data are not only different between different types of tissues and differences also exists between the cancerous and normal tissues. Theoretically, using the gene expression values is possible to detect all types of cancers accurately.

Gene expression data is new to the field of data mining and it gives challenges to those existing algorithms by the flood of gene expression data as well as their complexity. High dimensionality is the major challenge for the existing data mining algorithms in the knowledge discovery process. Inherited by the physical properties of the microarray experiments, the gene expression data consists of large number of numerical attributes, the gene expression values, but limited records. For example, there exist microarray experiments that can examine about 40,000 genes from 10 samples under 20 different conditions in one single experiment (Brazma 2000). The sparsity of data in the high-dimensional space, referred as dimensionality curse, causes the meaningfulness of proximity or clustering being questioned (Aggarwal 1999, 2000). Moreover, most of the available datasets contain very limited number of records compared with the number of available attributes. For example, the "Subtype of childhood leukemia" dataset (Li 2003) from St. Jude Children Research Hospital contains more than 12,000 genes but only 327 samples. It is very difficult for us to obtain a good approximation of the real world from the data. Last, but not the least, the readability and understandability are important to these biological related studies. It is because any hypothesis make by our knowledge discovery process requires biologists to conduct experiments for validation. In most of the cases,

bioinformatics technologies should be capable to provide good approximation of the real world from the limited number of high dimensional data and those techniques that give easy understandable results to the biologists will become the winners in the competitions.

Data mining techniques, such as pattern association and clustering, are now frequently applied in cancer and gene expressions correlation studies. They are expected to discover the cancer causing gene expression patterns for different diagnosis purposes, e.g., identifying the development of cancers in earlier stages and proposing useful treatment plans (Li 2002a). However, the usefulness of the gene expression data analysis obtained by traditional data mining techniques, such as clustering, are still questionable due to the aforementioned dimensionality curse problem (Aggarwal 1999,2002). Furthermore, clusters formed by conventional measures of tightness of data points often lack of practical biological meaningful support and they are not easy to understand and become applicable in the knowledge discovery process in biology domain. On the other hand, low occurrence patterns may still important in bioinformatics problems, especially in cancer detection that scientists are trying to find out abnormal gene expression patterns with relatively low occurrence. Most of the patterns mining algorithm cannot discover those low support patterns efficiently and the efficiency problems are become more serious in front of the high dimensionality of the gene expression data. In general, traditional data mining techniques are insufficient to provide meaningful and understanding analysis for the cancer detection problem efficiently.

Two newly introduced data mining techniques, projected clustering (Aggarwal 1999,2002) and emerging patterns (Dong 1999), are targeted for grouping high dimensional data and cancer detection respectively.

Projected clustering is targeted for grouping the high dimensional data, such as gene expression data, in lower dimensional subspaces in order to tackle the problem of dimensionality curse. It minimized the information loss in dimension reduction process by projecting those high dimensional data points into different lower dimensional subspaces for different clusters. Even the resulting clusters are more meaningful in lower dimensional subspaces, the distance measures used in projected dimension is not capable to capture any biological significant information. Thus, the resulting clusters may not be biologically meaningful, easy to understand for biologists and the may not applicable to separate the cancerous and normal tissues in the cancer detection problem.

Emerging patterns (EPs) have high discrimination power are designed for the classification problems. It can capture the biological significant information from the data. For example, the EPs only existed in cancerous tissues but never occur in normal tissues are potentially consists of sets of cancer causing genes. EPs are also easy to understand because they are just collections of attributes in dataset and this property is especially important for bioinformatics applications. In addition to the readability, there are some existing mining algorithms that can retrieve

low occurrence EPs from high dimensional data and it is important for cancer related studies since less frequently occurring patterns may be as important as frequently occurring patterns. However, the volume of EPs generated is very large for high dimensional gene expression data (Li 2001). In order to maintain the efficiency in using EPs, we have to use top EPs instead of full set of EPs in applications. Although the accuracy in classification is still high as reported in (Dong 1999b, Li 2002a), the robustness of EP-based classifier for the new data is still questionable. Increasing the number of top EPs used in applications can absolutely improve the robustness of the classifier but the efficiency will be suffered and such a strategy should not be an ultimate solution.

In this paper, we use the emerging pattern based projected clustering technique and the gene expression data in the problem of cancer detection. In the past, emerging patterns and projected clustering techniques were used independently in solving different types of problems since they are strong in different domains. In this paper, we propose to integrate them to form effective and yet easy-to-understand clusters of gene expression data. Moreover, the resulting clusters are used to classify the unseen data (cancerous and normal tissue) in the cancer detection problem. The classification performance of those projected clusters obtained by our EPPC algorithm and ORCLUS (Aggarwal 2002) are also compared. The main idea of our approach is to introduce the readability and strong discriminatory power of EPs in the dimension projection process of the projected clustering so that the readability of the projected clusters can be improved. The complete framework is illustrated in Figure 1. The rest of the paper is organized as follows. The use of EPs is briefly described in Section 2. In Section 3, we present our EP-based projected clustering (EPPC) approach. The dataset used and the experimental results are reported in Section 4. The final section concludes the paper.

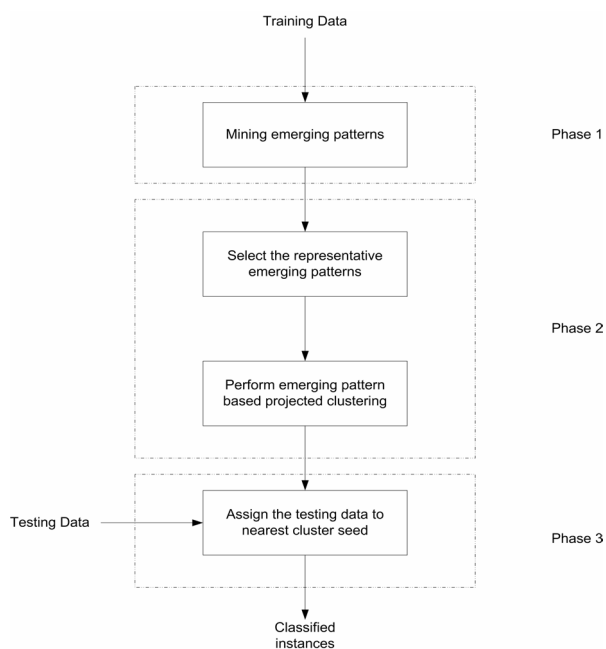


Figure 1: Flowchart of the EPPC framework

2 Use of Emerging Patterns

Dong and Li (1999) defined the emerging patterns as itemsets whose supports increase significantly, larger than the threshold value called the growth rate (ρ), from one dataset (D_1) to another (D_2). There are several types of EPs, such as Jumping EPs (JEPs), plateau EPs and so on. All of them have different properties and they are applied to different problems. They are easy to understand because they are only collections of data attributes. For example, a cancerous EP (Li 2001):

$$\{\text{gene(K03001)} \geq 89.20\} \text{ and } \{\text{gene(R76254)} \geq 127.16\} \text{ and } \{\text{gene(D31767)} \geq 63.03\}$$

is a pattern that only occurs in cancerous tissues but not in normal tissues.

Dong and Li (1999) also introduced border based mining algorithm, such as MDB-LLBORDER and BORDER-DIFF, to discover EPs. In classical patterns association analysis, those mining algorithms make use of mainly one nice property called subset-closedness of frequent itemsets to extract useful high occurrence patterns (Dong 1999). The most famous representative is perhaps the Apriori algorithm's candidate generation (Agrawal 1994). However, this category of algorithms cannot mine the low occurrence patterns efficiently since the candidate sets for low occurrence patterns needed to explore is huge. By using interval-closed properties (Dong 1999) of the itemsets, an itemset can be represented by an order pair called border (Dong 1999). With the help of the border based mining algorithm, such as MDB-LLBORDER (Dong 1999) and BORDER-DIFF (Dong 1999), the set of EPs is just a difference between high occurrence pattern borders in the two dataset partitions, i.e. the difference between two sets of frequent patterns. The interval-closed properties together with the border based mining algorithm solved the problem of extracting low occurrence EPs efficiently and a complete set of EPs can be retrieved. The algorithmic flow of EPs mining is illustrated in Figure 2 below.

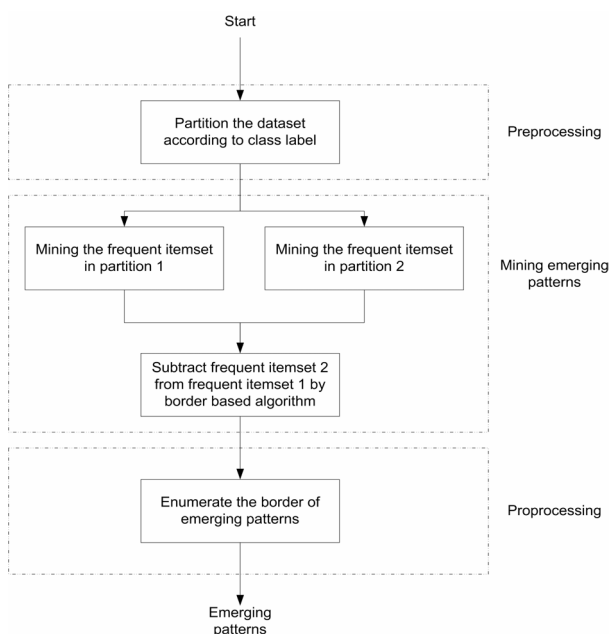


Figure 2: Flowchart of mining of EPs

EPs are previously used in cancer related classification problems. Since microarray experiments are still expensive in terms of time and cost, they are only being conducted to investigate some of the biological significant properties, the data itself always contains predefined class labels of important biological meanings. For example, in cancer related data, the data are always labelled as cancerous and normal. The existence of class labels in most of the cancer related dataset are necessary for EPs extractions and make EPs applicable to the cancer detection problem.

Another advantage of EPs is their discrimination power in classification. EPs capture the significant differences of attribute's values that exist in different partitions of the dataset. So, EPs are naturally identifying those collections of attributes of data points that are close in their values within the same cluster but far away to others. In terms of the discrimination power, the EPs with highest growth rates are most preferred. Therefore, the JEPs, whose growth rate is infinity, are being employed in our proposed method. However, the number of JEPs is still numerous (Li 2001). For the sake of simplicity and efficiency, we use the top 20 cancerous tissues JEPs and the top 20 normal tissues JEPs mined (Li 2002b) from 35 top-ranked genes by entropy method in this study.

3 EP-based Projected Clustering (EPPC) Algorithm

Before proceeding to describe our EP-based projected clustering algorithm, we introduce some notations and definitions. Let N be the total number of data points and n_i be the number of data points in cluster C_i . Assume that the dimensionality of full data space D is equal to d and the dimensionality of projected space D_i of cluster C_i is equal to d_i , where $d_i \leq d$. Let $X_i = \{\vec{x}_1, \vec{x}_2, \dots, \vec{x}_n\}$ be the set of data points in cluster C_i , $\vec{p}_j = \{x_{j1}, x_{j2}, \dots, x_{jd_i}\}$ be the projected point and \vec{s}_i be the centroid, i.e. $\vec{s}_i = \sum_{j=1}^{n_i} \vec{p}_j / n_i$. Finally, the projected distance of projected point \vec{p}_j to the center of cluster C_i is written as $Pdist(\vec{p}_j, \vec{s}_i, D_i)$.

Finding projected clusters can be established as a two-fold problem (Aggarwal 1999, 2002). First, we have to locate the clusters' center. Second, we needed to find out the projected dimensions for the corresponding clusters. In this paper, we focus on the dimension projection part of the problem.

The proposed EPPC algorithm includes three phases similar to (Aggarwal 2002), namely, initialization, iteration and refinement. In general, the initialization phase is to pick the initial cluster seeds for the iteration phase. In the iteration phase, data points are assigned to different clusters and the projected dimensions of those newly formed clusters are being evaluated. The iteration phase continues to improve the quality of clusters until the number of user specified clusters are obtained. Once the set of best cluster seeds is obtained after iterations, the refinement phase will start and all the data points will be

reassigned to those cluster seeds obtained by the iteration phase to form the final clusters. These three phases of the proposed EPPC algorithm are detailed in Figures 3 and 4.

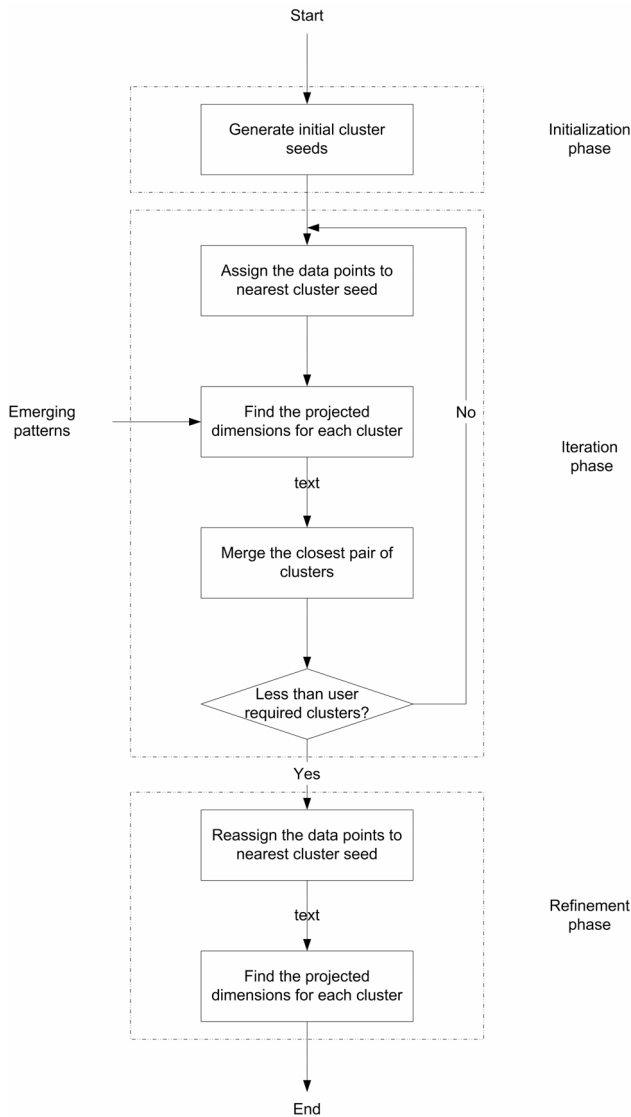


Figure 3: Flowchart of the EPPC algorithm

3.1 Initialization Phase

In this phase, the number of final clusters is defined by the users. We randomly pick k_0 initial cluster seeds from the dataset, where k_0 should be several times larger than k , and the projected dimensions of all initial seeds are initialized to the full dimensions of the dataset initially.

3.2 Iterative Phase

The goal of the iteration phase is to improve the quality of the cluster seeds iteratively in order to find the best clusters. There are three operations in this phase, namely, assignment, dimension projection and merging.

For the assignment operation, there should be k_c cluster seeds in the current iteration. In this operation, the data points in the dataset are assigned to their closest seed. We use the distance metric, such as city segmental distance or

Euclidean distance, to measure the distance between the data points and cluster seeds under those projected dimensions, i.e. the projected distance, $Pdist(\vec{p}_j, \vec{s}_i, D_i)$.

After the partitions are formed, the centroids of each partition are evaluated and they are used as the new seeds in the next iteration. This procedure is illustrated in Figure 5.

For the dimension projection operation, those partitions formed by the assignment operation consist of a set of data points. In this operation, the projected dimensions of each projected cluster are evaluated by their own data points. For each partition, we examine its data points and find those EPs embedded. The user specified numbers of EPs that are with most frequent occurrence are chosen and this set of EPs act as the collection of projected dimensions for that particular partition. This procedure is described in Figure 6.

In the last operation, i.e. the merging operation, the closest pair of clusters is merged together to form a new cluster. They are obtained by evaluating the average distance between the union of data points and new cluster seed of merged clusters. The smaller the average distance, the closer the pair of clusters. The details of this operation can be found in Figure 7.

3.3 Refinement Phase

Finally in the refinement phase, the resulting cluster seeds obtained from the iteration phase are then used to form the final clusters by assigning all the data points to them again. The goal of this phase is to ensure that all data points are assigned to the closest cluster seeds after the final cluster seeds are found.

```

Algorithm EPPC ( $k_0, k, E$ ) {
  Initialization phase
  Pick  $k_0 > k$  initial cluster seeds randomly from the dataset;
  Set no. of current cluster to no. of initial cluster;
  for each cluster {
    Set the cluster dimension to full dimensionality
  }
  Iterative phase
  While no. of current cluster > user requirement {
    Assign the data points to the nearest cluster seeds;
    Determine the cluster dimensions associated to each cluster;
    Merge the closest clusters and obtain the new seed for the newly merged cluster;
    Update the no. of current cluster;
  }
  Refinement phase
  Reassign the data points to the set of good seeds obtained from iteration phase;
  Determine the cluster dimensions associated to each cluster;
  Return the projected clusters with cluster seeds, corresponding dimensions and data points;
}
  
```

Figure 4: The EPPC algorithm

```

Algorithm Data_Point_Assignment {
  for each data point {
    for each cluster {
      Determine the projected distance between
      the data points and current seeds;
    }
    Add the data points to their nearest cluster;
  }
  Remove cluster from the set if it is empty;
  Set the centroids of those projected clusters as the new
  cluster seed;
  Return the cluster seed and data set in projected
  clusters;
}

```

Figure 5: Data point assignment algorithm

```

Algorithm Dimension_Projection {
  for each cluster {
    Find the user specified number of EPs that having
    most frequent occurrence among the data points
    in the cluster;
    Find the corresponding attributes that make up the
    corresponding set of EPs;
    Set the projected dimensions to that collections of
    attributes;
  }
  Return the dimensions for the projected clusters;
}

```

Figure 6: Dimension projection algorithm

```

Algorithm Cluster_Merging {
  for each pair of cluster {
    Find the closest pair of clusters from the set of
    existing clusters;
    Merge the data points of the two clusters;
    Find the projected dimensions of the unified data
    points;
    Find the new seed of the unified data points;
    Evaluate the radius of the merged clusters;
  }
  Merge the closest pair of clusters such that the
  radius of the merged clusters is minimal;
  Return the set of new cluster seeds;
}

```

Figure 7: Cluster merging algorithm

4 Experimental Results

In order to studies the use of gene expression data and emerging pattern based projected clustering for the cancer detection problem. We implemented EPPC framework with MATLAB and used colon tumor dataset from Alon (1999) for cancer detection experiment. We also implemented the ORCLUS with MATLAB and experimented it by using the same dataset. The experiment performance of the cancer detection in terms of the accuracy in identifying the correct tissue types and the readability of resulting clusters are illustrated below.

4.1 Dataset

The colon tumor dataset was collected by Alon et al at Princeton University. The dataset consists of 2000 gene expression values of 40 tumor and 22 normal colon tissues samples and it is publicly available at <http://microarray.princeton.edu/oncology/affydata/index.html>. Since the original dataset consists of 2000 gene expression values as attributes for total 62 samples and not all of those attributes are useful in separating samples into different classes, the dataset is first reduced its size by using the entropy discretization method provided by MLC++ (Kohavi 1994). The entropy method finds a total of 135 significant genes that relevant to classify those samples and we pick the 35 top-ranked genes as mentioned in (Li 2002b) to generate a reduced dataset. In order to facilitate the study of the relationship between different gene expression values and the types of tissue, each gene expression value of the reduced dataset is normalized before they undergo the clustering process. The mean of the normalized gene expression values is equal to zero while the standard deviation is equal to one.

4.2 Experimental Setting

In this paper, we report and compare the classification performances of those projected clusters generated by our EPPC algorithm and ORCLUS (Aggarwal 2002). Initially, a specified portion of tumor and normal samples are randomly selected and used as the training dataset to generate projected clusters. Then, the rest of the records act as testing samples and they are assigned to those resulting projected clusters accordingly. Measurement of the classification accuracy is defined below.

$$\text{Classification accuracy} = \frac{\text{No. of testing samples in correct clusters}}{\text{total number of testing samples}}$$

In order to minimize the effect of initial points and ordering problems in clustering, we repeated every experiment for 50 times for different number of initial clusters and final clusters, with training and testing samples that selected and ordered randomly.

4.3 Classification Performance

4.3.1 ORCLUS project clusters

In this experiment, we studied the classification performance of ORCLUS projected clusters. The focus of this experiment is on the effects of ORCLUS projected clusters with different number of projected dimensions and their performance in classification for the colon dataset (Alon 1999). We implemented ORCLUS and used Euclidean distance as a distance metrics by following the methodology in (Aggarwal 2002). Each projected dimension is equivalent to one principle component of the covariance matrix of the datasets. The details of the dimension projection mechanism are available in (Aggarwal 2002). In this test, we use 70% of samples (43 samples with 28 tumor tissues and 15 normal tissues) from the dataset as training data and choose the number of principle components (I) that we are interested in to 1, 5, 10, 15, 16, 17, 18, 19, 20 and 35. The averaged classification accuracy of those 50 repetitions with respect

to different number of initial clusters and final clusters are studied.

From the experiment results, we found that the relationship between the classification accuracy and number of projected dimensions are similar among different of number of initial clusters. For example, Figure 8 and Figure 9 are the summaries of the experimental results in different number of initial clusters and they give the similar tendency in classification performances.

Figure 8 summarized the experiment results with the number of initial clusters equal to 40 is employed to demonstrate the variation in classification performances of projected clusters with different number of projected dimensions. We found that if we used less than 15 principle components as projected dimensions to form ORCLUS projected clusters. The resulting clusters give relatively low classification rate for those testing samples in cancer detection. By increasing the number of the projected dimensions from 1 to 15, the classification accuracy of those projected clusters are increased significantly, as shown in the Figure 8. In the case of more than 15 principle components are used, the classification rate is much higher in general but the improvements in terms of the classification accuracy by further increasing the number of projected dimensions are flattened. The reason is that the increase the number of projected dimensions will increase the volume of information extracted from the training data in order to form more reliable clusters, thus the higher classification rates are obtained in general when the number of projected dimensions are increased. But if the number of projected dimensions is reached the optimal, any further increment of projected dimensions will not introduce relevant information. Besides, it will introduce additional distance between the testing samples and the projected clusters' centers and it may cause the distance between the testing samples and those cluster centers become too close to distinguish from the most desirable cluster and affected the classification accuracy. It is referred as dimensionality curse problem in literature (Aggarwal 2002).

We can deduce the optimal number of projected dimensions with respect to different number of final clusters from Figure. 8. According to our experiment results, if the number of final clusters is smaller than 24, the projected clusters formed by 35 principle components give the highest classification rate. However, projected clusters generated by 20 principle components give the highest classification accuracy when the number of final clusters are larger than 24. It is explainable that the smaller in the number of the final clusters, cluster centers are far away to each other. The distance between those testing samples and different resulting projected cluster centers are always in greater differences. Therefore, the higher dimensionality may not degrade the classification accuracy very much and the additional dimensions may give more relevant information for the testing samples to distinguish the correct clusters from the rest of undesired clusters. That is the reason for those projected clusters with higher dimensionality classified those testing samples better if the number of final clusters is small. If the number of projected clusters increased, the distance

between them decreased. The negative effects of increasing dimensionality of the projected clusters overwrite its benefits and then the drop in performance occurred. That is the reason why projected clusters with 20 dimensions give higher classification rate when compare with projected clusters with 35 dimensions if the number of clusters larger than 24 in Figure 8.

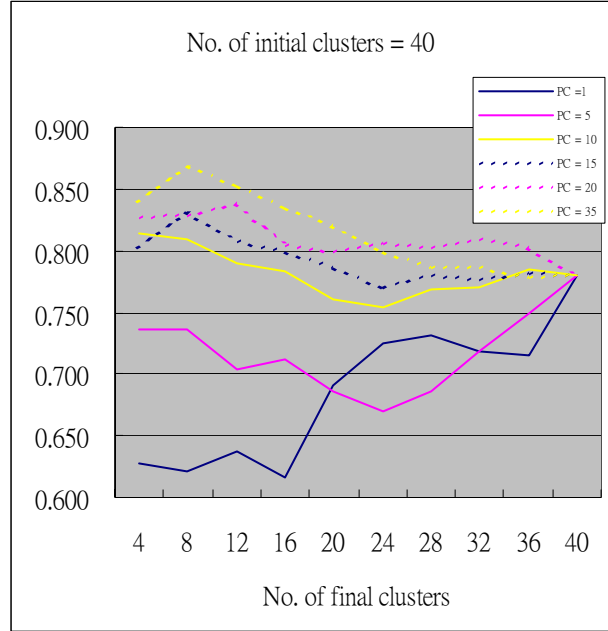


Figure 8: Classification accuracy of ORCLUS projected clusters (Initial cluster = 40)

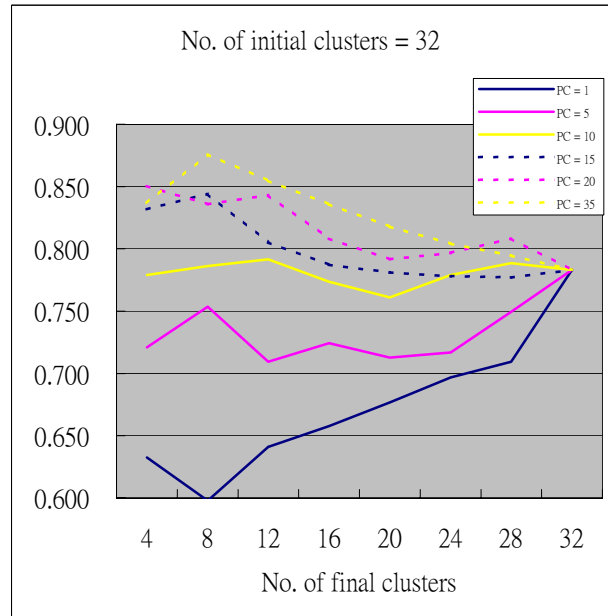


Figure 9: Classification accuracy of ORCLUS projected clusters (Initial cluster = 32)

Since our reduced dataset consists of 35 dimensions, the projected clusters with 35 dimensions do not provide any advantages in dimensional reduction aspect and it may not very interesting for further studies. However, the classification accuracy of projected clusters with 20 and 15 projected dimensions is very closed to those projected

clusters with 35 projected dimensions as shown in Figure 8 and they are more applicable in real applications. In Figure 10, we provide a detail investigation to the classification performances of those projected clusters with 15 to 20 dimensions. In terms of classification accuracy, we cannot find an optimal number of dimensions easily from Figure 10 that can outperform the others in cancer detection. Since the differences in classification rate between those projected clusters as shown in Figure 10 is very small. We use the classification rate of projected clusters consists of 17 projected dimensions as representative, as shown in Table 1, for comparison purpose in the rest of the paper, because its performance is generally good in different number of final clusters.

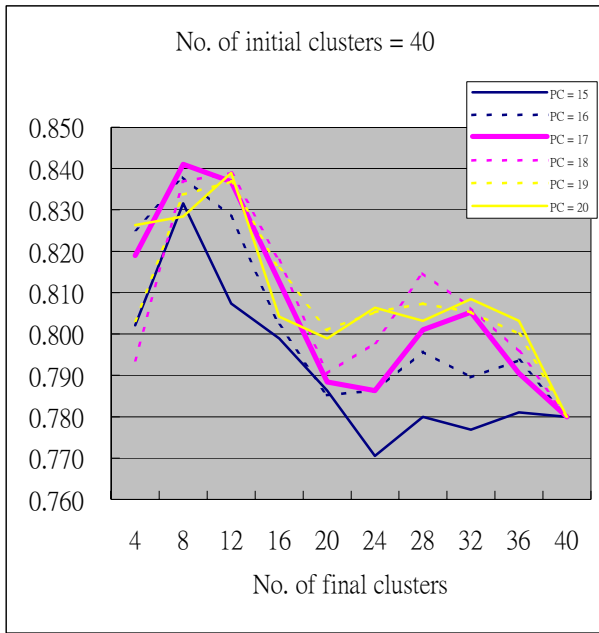


Figure 10: Classification accuracy of ORCLUS projected clusters (Initial cluster = 40)

Initial Cluster Num.	Final Cluster Num.									
	4	8	12	16	20	24	28	32	36	40
8	0.846	0.838								
16	0.844	0.841	0.819	0.809						
24	0.833	0.845	0.843	0.807	0.815	0.796				
32	0.829	0.833	0.815	0.814	0.799	0.799	0.804	0.783		
40	0.819	0.841	0.837	0.813	0.788	0.786	0.801	0.805	0.791	0.780

Table 1: Classification accuracy of ORCLUS projected clusters (No. of PC = 17; 70% training data)

4.3.2 EPPC projected clusters

In this experiment, we study the effect of using different number of EPs to generate projected clusters by our proposed EPPC algorithm on the performance of the classification. We also used Euclidean distance as a metrics for the EPPC to obtain projected clusters. In this test, we employ the same set of training and testing samples that we used in previous experiment, i.e. 70% of samples from the dataset are selected and ordered randomly for training. Sets of EPs, consists of total 1, 3, 5,

6, 7, 8 and 10 EPs, are used to generate projection dimensions for the projected clusters in different sets of experiments and the average classification accuracy of those 50 repetitions are studied with different number of initial clusters and final clusters.

According to the experimental results, the relationship between the classification rate and the number of EPs used for dimension projection in EPPC is similar among different number of initial clusters and we used the Figure. 11 with 40 initial clusters for illustration. Figure 11 shows the classification accuracy of emerging pattern based projected clusters (n-EPPCs) and we found that if only one EP is used to generate the projected clusters, the classification rate such 1-EPPCs are not good enough when compared with 3-EPPCs and other n-EPPCs ($1 < n \leq 10$) in any number of final clusters. The major reason is that a single EP is unlikely to be adequate to capture all the significant dimensions that should be included in desired projected clusters. In most of the cases, using set of EPs to generate projected clusters can include more relevant and significant attributes and thus they always obtain higher classification rate as shown in our results. In general, more EPs used in generating projected clusters, the higher the classification accuracy can be obtained. In Figure 11, we can observe this trend by considering the improvement of classification performance from 1-EPPCs to 5-EPPCs. However, the situation becomes messy when the number of EPs used is more than 5. 5-EPPCs, 7-EPPCs and 10-EPPCs give the optimal classification accuracy in different number of final clusters but when the number of final clusters is smaller than 24, 10-EPPCs give better results and 5-EPPCs give the best classification rate when the number of final cluster is larger than 24. These findings are very similar to the results obtained in the previous set of experiments about the ORCLUS projected clusters studies. That is the limited number of final clusters mean the distance between them are larger and clusters with higher dimensionality may give results that better than the lower dimensional clusters. In these studies, they proved that the dimension projection by using EPs is reasonable since the experimental results obtained by n-EPPCs can be explained thoroughly by principles of clusters' dimensionality.

It is interesting that the classification rates obtained by 1-EPPCs are especially low when the number of final clusters either very small or very large. Its shape in Figure 11 shows that it is totally different from the others n-EPPCs. The reason behind is that if the number of final clusters are limited, the set of projected clusters that generated by single EP are not enough to capture the real world since single EP can only provide limited information for the projected dimensions. On the other hand, if there are too many final clusters, some of clusters' projected dimensions will be very similar because similar EPs are likely to be employed in the generation of projected dimensions and those similar clusters with relatively smaller inter-cluster distance may not distinguish the samples accurately.

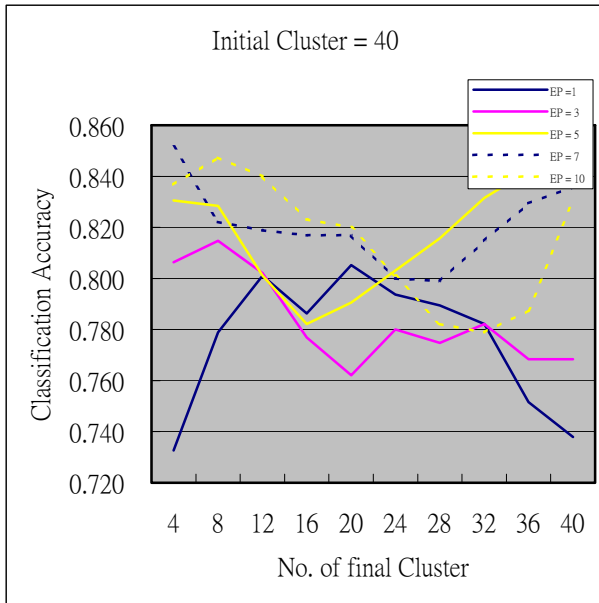


Figure 11: Classification accuracy of EPPC projected clusters (Initial cluster = 40)

From Figure 11, we found that the projected clusters generated by set of EPs with a number of used EP larger than 5 give very impressive classification rates in different number of final cluster. In Figure 12, we take a closer look to the classification accuracy from 5-EPPCs to 10-EPPCs. 5-EPPCs give the best classification result when the number of final cluster is larger than 24 but it does not perform well if the number of final cluster is less than 24. 10-EP project clusters perform similarly in opposite manner. However, 7-EP projected clusters perform well in all combination of final clusters generally and it can be used as the representative in comparative studies between ORCLUS projected clusters and n-EPPCs. The performances of 7-EPPCs in different environments are summarized in Table 2.

Initial Cluster Num	Final Cluster Num.									
	4	8	12	16	20	24	28	32	36	40
8	0.852	0.838								
16	0.839	0.823	0.807	0.807						
24	0.857	0.834	0.815	0.804	0.798	0.807				
32	0.844	0.829	0.822	0.826	0.816	0.802	0.803	0.823		
40	0.852	0.822	0.819	0.817	0.817	0.800	0.799	0.815	0.829	0.836

Table 2: Classification accuracy of 7-EPPC projected clusters (70% training data)

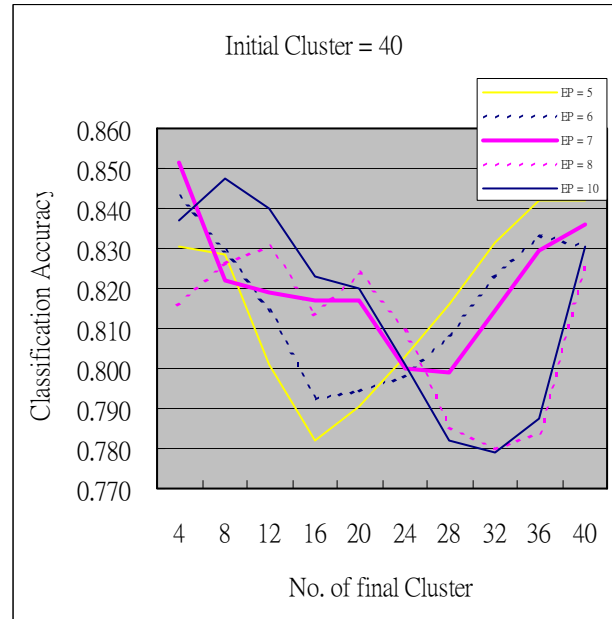


Figure 12: Classification accuracy of EPPC projected clusters (Initial cluster = 40)

4.3.3 Comparative studies on ORCLUS and EPPC

In above experiments, we have examined the performances of ORCLUS projected clusters and n-EPPCs in the cancer detection problem. Moreover, we have selected one representative from both technologies that has generally good performance in most of the examined conditions for comparison. They are the projected clusters with 17 principle components as projected dimensions obtained by ORCLUS and projected clusters with projected dimensions generated by set of seven EPs with EPPC.

Their performances in classification respected to different combinations of initial and final clusters number are compared and shown in the Table 2. The bolded entries in Table 2 are the experimental results that 7-EPPCs give a better performance when compare with ORCLUS projected clusters with 17 projected dimensions in Table 1. We found that the performance of 7-EPPCs obtained from our proposed EPPC algorithm is slightly better than the representative ORCLUS projected clusters. We got 16 conditions out of 30 that give better classification accuracy.

In addition to the classification power, other differences between ORCLUS and EPPC are summarized in Table 3. The major difference of EPPC is that it utilized the information in predefined classes, domain knowledge, which always available in gene expression data but ORCLUS do not make use of it. In the readability aspect, the result clusters' projected dimension of EPPCs are more easy to interpret. Further discuss on readability are available in later section.

	ORCLUS	EPPC
Use of class label (domain knowledge)	No	Yes
User inputs	No. of final cluster No. of principle component for dimension projection	No. of final cluster No. of EPs for dimension projection
Dimension projection	Using principle component	Using EPs
Individual set of dimensions for each cluster	Yes	Yes
Projected dimension	Linear combination of all existing attributes	Collection of attributes
No. of projected dimension for each cluster	Same in each cluster	Vary in different clusters
Classification accuracy	High	High
Readability	Bad	Good

Table 3: Summary of ORCLUS and EPPC

4.4 Readability – The projected dimension of resulting projected clusters

In traditional clustering algorithms, the resulting clusters only provide information on those data points that are said to be similar under a predefined distance metric. Most of the distance metrics, such as Euclidean distance, is geometrically meaningful with respect to the full dimensional data space. It is questionable to group data points in meaningful clusters (Aggarwal 1999,2002) and it would be very difficult for users to interpret when the number of dimensions of data is large.

Projected clustering works on a step further towards providing flexibility to individual clusters in having their own set of dimensions that they are significant to group the data points. More meaningful clusters can be formed in different subspaces instead of using the full data space and it successfully tackled the dimensionality curse problem. ORCLUS use the principle components generated from the covariance matrix of data as the projected dimensions of its projected clusters. In our first experiment, we show that ORCLUS is powerful to form projected clusters and those resulting clusters are applicable in the cancer detection (classification) with high classification accuracy. However, those projected dimensions in terms of principle components are not easy to interpret by users, especially for the biologists. A sample of tumor cluster with 5 projected dimensions is shown in Table 4. Each dimension, column in Table 4, of the clusters can be interpreted as a linear combination of 35 original dimensions in reduced dataset and they are never easy for user to interpret.

Our EPPC algorithms formulate projected clusters by using the discrimination power of the emerging patterns. In previous experiments, we showed that EPPC can form reliable projected clusters and those resulting clusters are also applicable in cancer detection problem. The classification accuracy of EPPC is comparable or even better than ORCLUS in some situations. More important is that the projected clusters formed by EPPC are just collections of attributes. They are easy to interpret by the

users. Three sample clusters are extracted and shown in Table 5 to demonstrate their readability. In this example, cluster 2 consists of 6 cancerous tissues that those tissues samples are similar in gene expression values with respect to two suspecting gene M76378 and T47377.

Gene No.	D1	D2	D3	D4	D5
M26383	0.097	0.239	-0.062	-0.030	0.051
M63391	-0.224	0.013	-0.277	-0.026	-0.074
R87126	0.163	-0.142	0.094	-0.017	0.300
M76378	-0.202	-0.053	0.551	0.064	-0.188
H08393	0.037	0.015	-0.040	0.002	0.025
X12671	0.077	0.034	0.011	-0.013	-0.029
R36977	-0.021	-0.051	-0.023	-0.017	-0.009
J02854	-0.011	0.271	0.065	-0.111	0.270
M22382	-0.012	-0.056	-0.114	0.010	-0.020
J05032	0.008	0.090	0.170	-0.013	-0.022
M76378	-0.157	0.040	0.029	0.068	-0.105
M76378	-0.055	0.052	-0.341	0.088	-0.090
M16937	-0.034	0.073	0.114	-0.018	-0.088
H40095	-0.012	-0.027	-0.309	0.003	-0.031
U30825	0.107	0.073	0.103	-0.080	0.090
H43887	-0.112	-0.084	-0.124	-0.046	-0.169
X63629	0.036	-0.030	0.162	0.029	-0.114
H23544	0.195	-0.084	-0.176	-0.027	-0.167
R10066	-0.060	-0.085	0.145	0.011	-0.077
T96873	0.045	-0.018	0.268	-0.033	0.035
T57619	0.001	0.253	-0.008	-0.034	0.015
R84411	0.147	0.110	0.003	-0.035	-0.254
U21090	-0.231	-0.336	-0.272	-0.083	0.425
U32519	-0.238	-0.396	0.086	-0.031	0.026
T71025	0.076	-0.120	0.167	-0.063	0.461
T92451	-0.169	-0.140	0.044	-0.052	-0.239
U09564	0.422	0.229	-0.048	0.024	0.067
H40560	0.177	-0.245	-0.155	-0.021	-0.055
T47377	-0.123	0.192	-0.007	-0.042	0.209
X53586	-0.064	0.013	0.047	-0.179	0.125
U25138	0.340	-0.070	-0.076	0.011	-0.192
T60155	0.178	0.119	0.009	-0.057	0.167
H55758	-0.024	-0.030	-0.012	-0.080	-0.092
Z50753	-0.474	0.493	-0.106	0.016	0.028
U09587	0.001	-0.011	0.000	0.946	0.145

Table 4: Sample tumor cluster – 5 dimensions projected cluster generated by ORCLUS

	No. of sample	Tissue Type	Cluster Dimension				
			Gene 1	Gene 2	Gene 3	Gene 4	Gene 5
Cluster 1	7	Normal	H51015	R10066	U32519	T47377	Z50753
Cluster 2	6	Cancer	M76378	T47377			
Cluster 3	10	Cancer	H08393	M76378			

Table 5: Sample cluster obtained by EPPC

5 Conclusions

In this paper, we propose an EPPC algorithm to generate projected clusters with sets of EPs. We also apply the projected clusters in the cancer detection problem and shows that the sets of EPs generate more reliable projected clusters and yield a higher classification rate than a single EP does. We performed a comparative studies on those projected clusters generated by ORCLUS and EPPC in different situations. The experimental results not only demonstrate the reliability and usefulness of using emerging patterns in dimension projection of projected clustering, but also show that the resulting clusters are more readable to end users and this feature is especially important to many bioinformatics applications. In our future work, we will examine how to generate optimal clusters in different situation by introducing new learning algorithm to select the EPs for dimension projection.

6 References

- Aggarwal, C.C., Procopiuc, C., Wolf, J.L., Yu, P.S. and Jong S.P. (1999): Fast algorithms for projected clustering. *Proc. of ACM SIGMOD International Conference on Management of Data*, 61-72.
- Aggrawal, R. and Srikant, R. (1994): Fast algorithms for mining association rules. *Proc. of Int. Conf. on Very Large Data Bases (VLDB'94)*, Santiago, Chile, 487-499.
- Alon, U., Barkai, N., Notterman, D.A., Gish, K., Ybarra, S., Mack, D. and Levine, A.J. (1999): Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. of the National Academy of Sciences*, **96**(10):6745-6750.
- Dong, G. and Li, J. (1999a): Efficient mining of emerging patterns: discovering trends and differences. *Proc. of the fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Diego, California, 43-53.
- Dong, G., Zhang, X., Wong, L. and Li, J. (1999b): CAEP: classification by aggregating emerging patterns. *Proc. of Second International Conference on Discovery Science (Lecture Notes in Artificial Intelligence Vol.1721)*, 30-42, Springer-Verlag.
- Kohavi, R., John, G., Long, R., Manley, D. and Pfleger, K. (1994): MLC++: A Machine Learning Library in C++. In *Tools with Artificial Intelligence*. 740-743.
- Li, J and Wong, L. (2001): Emerging patterns and gene expression data. *Proceedings of 12th Workshop on Genome Informatics*, , Tokyo, Japan, 3-13.
- NCI: Understand CGAP, National Cancer Institute.: <http://press2.nci.nih.gov/sciencebehind/cgap/cgap01.htm>. Accessed 22 Apr 2003.
- NCI: Understanding Cancer, National Cancer Institute.: <http://press2.nci.nih.gov/sciencebehind/cancer/cancer01.htm>. Accessed 22 Apr 2003
- Aggarwal, C.C. and Yu, P.S. (2002): Redefining clustering for high-dimensional applications. *IEEE Trans. on Knowledge and Data Engineering* **14**(2):210-225.
- Brazma, A., Robinson, A., Cameron, G. and Ashburner, M. (2000): One-stop shop for microarray data - Is a universal, public DNA-microarray database a realistic goal? *Nature* 699-70.
- Li, J., and Wong, L. (2002a): Identifying good diagnostic gene groups from gene expression profiles using the concept of emerging patterns. *Bioinformatics* **18**(5):725-734.
- Li, J and Wong, L.: (2002b): Identifying good diagnostic gene groups from gene expression profiles using the concept of emerging patterns (Corrigendum). *Bioinformatics* **18**(10):1406-1407.
- Li, J., Liu, H., Downing, J.R., Yeoh, A.E. and Wong, L. (2003): Simple rules underlying gene expression profiles of more than six subtypes of acute lymphoblastic leukemia (ALL) patients. *Bioinformatics* **19**(1):71-78.