Principal Directon Divising Partitioning initialisation of K-means Clustering allows to identify the most salient genes in discriminating among Leukemias

Diego Liberati

National Research Council of Italy

Abstract

This paper attempts to cluster leukemia patients described by gene expression data, and to discover the most discriminating genes that are responsible for the clustering. A combined approach of Principal Direction Divisive Partitioning and bisect K-means algorithms is applied to the clustering of the investigated leukemia dataset. Both unsupervised and supervised methods are considered in order to get optimal result. The combination of PDDP and bisect K-means successfully clusters leukemia patients, and efficiently discovers salient genes able to the discriminate the clusters. The combined approach works well on the automatic clustering of leukemia patients depending merely on the gene expression information, and it has great potential on solving similar problems, like classifying pancreatic tumors. The salient identified genes may thus enhance relevant information for discriminating among leukemias.

1. Introduction

The rapid development of the DNA micro-array technology is making it more and more convenient to obtain various gene expression datasets with abundant information that can be very helpful for many meaningful biomedical applications such as prediction, prevention, diagnosis and treatment of diseases, development of new drugs, patient-tailored therapy, precision and personalized medicine. However, these datasets are usually very large and unbalanced, with the number of genes (thousands upon thousands) being much greater than the number of patients (generally from tens to hundreds). Consequently, how to analyze effectively this kind of large datasets with few samples and numerous attributes, for example, how to classify according to their gene expression profile the patients suffering from certain disease, or how to determine from thousands of genes the most discriminating ones that are responsible for the corresponding disease, should be viewed as an important issue.

In the recent decades there have been many exciting research results in the area of DNA micro-array data mining on the basis of gene expression data analysis. For instance, to cite a few pioneer results, depending solely on gene expression monitoring to micro-array datasets, Golub et al (1999) classified sample patients of acute leukemia as two sub types, ALL (Acute Lymphoblastic Leukemia) and AML (Acute Myeloid Leukemia), and predicted the sub types of new leukemia cases according to the expression values of the most decisive genes that were discovered during the classification of sample cases; Scott et al (2002) discovered a new sub type of acute leukemia, MLL (Mixed Lineage Leukemia), claimed as distinct enough to be separated from ALL or AML; In a hierarchical point of view, Loris et al (2004) classified patients of advanced ovarian cancer and extracted significant genes which characterized each level in the hierarchies; On the basis of gene expression profile analysis van’t Veer et al (2002) predicted the clinical outcome (relapse / non-relapse) of breast cancer and Pomeroy et al (2002) predicted the outcome (survivor / failure) of embryonal tumor of central nervous system; Alon et al (1999) clustered correlated gene families about colon tissues and separated cancerous from non cancerous tissues; Dinesh et al (2002) performed the tumor versus normal classification of prostate cancer and predicted the clinical outcome of prostatectomy; Eng-Juh et al (2002) classified the sub types and predicted the outcome of pediatric acute lymphoblastic leukemia; Gavin et al (2002) separated malignant pleural mesothelioma (MPM), which is not a lung cancer, from adenocarcinoma (ADCA) of the lung; Alizadeh et al (2000) identified two distinct types of diffuse large B-cell lymphoma (DLBCL), the germinal centre B-like DLBCL and the activated B-like DLBCL.

The technologies applied in the analysis of gene expression data are various. In Golub et al. (1999) a method of neighborhood analysis is used to select out the most informative genes that are related to the classification of patients, a class predictor is designed by using the sum of the weighted votes from these genes to determine the wining class, and a cross-validation method is adopted to test the accuracy of the predictor. To classify the leukemia patients, a technology of self-organizing maps is applied to obtain two classes. In van't Veer et al. (2002) an unsupervised method is used to cluster both genes and tumors, and a supervised alternative is adopted to identify the outcome of the tumors and extract the most significant genes that are related to the outcome. In Pomeroy et al. (2002) Principal Component Analysis (PCA) is applied to determine different types of tumors and the related genes. In Alon et al.(1999) a deterministic-annealing algorithm is used to organize both genes and sample tissues into binary trees so that they can be clustered hierarchically. In Gavin et al. (2002) gene expression ratios are calculated and thresholds are selected to distinguish between cancer and non-cancer tissues.

In this paper, an approach based on the collaboration of three algorithms, Principal Component Analysis (PCA), Principal Direction Divisive Partitioning (PDDP), and bisect K-means, is applied to cluster the sample patients from a public leukemia dataset (Scott et al., 2002) consisting of 72 leukemia samples (24 acute lymphoblastic leukemia (ALL), 20 mixed-lineage leukemia (MLL) and 28 acute myeloid leukemia (AML), each sample being represented by 12,582 gene expression values. In the mean time, the few significant genes more determinant to the clustering results are identifieded.

The rest of the paper is organized as follows:

Section 2 is about the description and the pre-processing of the leukemia dataset that is used in the experiments, Section 3 reports the salient issues about the clustering algorithms, Section 4 illustrates the experimental results of clustering the leukemia dataset, Section 5 discusses the results.

2. Dataset Description and Pre-processing

2.1 Description of the Dataset

The dataset analyzed in this paper is the combination of two leukemia datasets processed in Scott et al. (2002), where 57 samples (20 ALL, 17 MLL and 20 AML) are used for training and 15 (4 ALL, 3 MLL and 8 AML) for testing the clustering of leukemia patients. Each patient is determined by a sequence of 12,582 real numbers, each measuring the relative expression of the corresponding gene. The data set can then be viewed as 72 points in a 12,582-dimensional Euclidean space. A simple measure of the genomic difference between two patients can be obtained by resorting to the Euclidean distance of two points. In order to ease the algebraic manipulations of data, the dataset can also be represented as a real 2-D matrix S of size 72×12,582; the entry s_{ij} of S measures the expression of the j^{th} gene of the i^{th} patient.

2.2 Pre-processing of the Dataset

The leukaemia dataset is a very large matrix with more than ten thousand genes as its columns, while a great portion of them, with small changes of values between different patients, provides much less information related to the patient clustering than the residual small portion, in which large differences of values can be found between different patients or patient types. In this dataset, it can be observed that a very large portion of genes has relatively small standard deviation values, although the values vary from 0 to 15,000. For example, at least 10,000 standard deviation values are less than 1,200. Therefore, prior to the patient clustering, it is possible to apply a filter to remove those genes of little importance (Garatti et al., 2007). In order to analyze such a huge dataset without any filters, a higher amount of time and storage would be needed, as well as a larger amount of computational resources. The removing of less important genes can help decrease the complexity of analysis and the requirement of computational resources without much affecting result precision. Furthermore, the removing of those genes may also reduce the interference caused by noise.

By taking all these factors into account, a pre-processing of the dataset is applied first to remove those genes with small standard deviation values. A threshold 400 is used to filter out the genes with standard deviation values less than it. The dataset after this pre-processing almost halves, becoming a 72×6,611 matrix with the removing of 5,971 gene columns. The reason for using 400 as the threshold is that it keeps a large portion of the data, so that the important information will reasonably not be ignored, at the same time removing another large portion– almost a half - of data to speed up the clustering procedures. In the following sections, unless otherwise specified, all the analysis is based on the 72×6,611 dataset after the pre-processing with threshold th = 400.

3. Description of Algorithms

The clustering analysis of the leukemia dataset is based on three steps. First, with the principal component analysis, all the genes in the dataset are sorted according to their significance to the patient clustering. Then, the dataset is clustered using a modified bisect K-means algorithm which is essentially the combination of principal direction divisive partitioning, using to initialize the following, and K-means. Finally, the minimum set of genes minimizing clustering errors is identified. This gene set consists of a few necessary and sufficient genes in the sense of the clustering approach applied in this paper, but the so identified genes are also keen to provide useful information for the differential diagnosis and even better understanding of the corresponding sub types of leukemia.

3.1 Principal Component Analysis (PCA)

It is well known that the PCA method (Hand et al.,2001) (O'Connel 1974) (Wall et al. 2003) works better on measuring the contribution of attributes of samples to the clustering when the dataset can be linearly partitioned. The extraction of principal components is briefly recalled as follows for the sake of completeness:

Given a p×N dataset S where p and N are respectively the numbers of samples and attributes. If dataset S is a centralized matrix where each column (i.e. attribute) of S has zero mean value, then the first principal component of S should be the eigenvector corresponding to the largest eigenvalue of the covariance matrix of S, namely $S^T S$, the second principal component of S should be the eigenvector corresponding to the second largest eigenvalue of $S^T S$, and so on. A simple proof is given out in (Hand at al., 2001).

The principal components can be obtained from the singular value decomposition (SVD) (Wall et al., 2003) of S as the product of three special matrices: the orthonormal unitary square matrix $U_{P\times P}$ (i.e. $U^{-1}=U^T$), the diagonal matrix $\Sigma_{P\times N}$, and the orthonormal unitary square matrix $V_{N\times N}$ (i.e. $V^{-1}=V^T$). Any non-zero diagonal element of matrix Σ is called a singular value of matrix S (i.e. the square root of an eigenvalue of matrix $S^T S$), and the columns of matrix V (i.e. the eigenvectors of $S^T S$) corresponding to the largest singular values are in turn the principal components of S.

When a principal component, generally the one corresponding to the largest singular value, is selected out, the degree of contribution of the attributes to the clustering of samples can be quantified by comparing the absolute values of the elements in the principal component vector. The positions of the largest absolute values point out the most discriminating attributes for clustering the sample.

When the dataset matrix S is not centralized, with the mean values of some attributes being non-zeros, the SVD should be performed on the centralized form of S so as to equally weight the contribution from each attribute.

3.2 Principal Direction Divisive Partitioning (PDDP)

The PDDP algorithm is proposed by Boley (1998). It has the following steps:

(1) For the matrix S (in general S is not centralized) in Section 3.1, first calculate

the mean value vector w=[w_1, w_2, …, w_N] for all the samples. The mean value vector is the centroid of the samples, where $w_j = \frac{1}{p}\sum_{i=1}^{p} s_{ij}$ (1≤j≤N) and s_{ij} is the element in the i^{th} row and j^{th} column of S.

(2) Calculate matrix S_0, the centralized form of S, as S_0 = S - ew and e = $\overbrace{[1,1,\cdots,1]}^{p}{}^{T}$. Then, by the PCA analysis described in Section 3.1, decompose S_0 as S_0 = UΣV.

(3) Select an appropriate principal component v = $[v_1,v_2,\ldots,v_N]^T$ for S_0 , where vector v is determined either manually or automatically by the method described in Section 3.3.

(4) Write matrix S as $[S_1,S_2,\ldots,S_p]^T$. If $(S_i\text{-}w)v\leq 0$, then $S_i \in S_L$, otherwise $S_i \in S_R$, where 1≤i≤p.

The rationale of PDDP has a geometrical interpretation. The p×N dataset is first transformed to an N-dimensional coordinates system originting at the dataset centroid and having all the N component vectors (even not principal) as coordinates. Suppose a principal component is selected to do PDDP, then the data points are separated as two clusters by an (N-1)-dimensional hyperplane passing through the origin and is normal to such principal component vector. Generally speaking, some distance based methods - such as the minimum distance and the average distance between two different clusters - can be used to measure the difference between them.

It should be pointed out that PDDP can be applied repeatedly to any cluster to get two sub clusters; therefore any number of clusters can be obtained by iteratively using such algorithm. Savaresi et al. (2002) have proposed a method to tell which one of two given clusters is more suitable to be further split, while Kruengkrai et al. (2003) have suggested how to determine wether a cluster culd again be split, thus helping to terminate iterations.

3.3 The Selection of Principal Components

3.3.1 A possible problem of principal component selection

The selection of an appropriate principal component is the precondition of the success of PDDP clustering. In general, the first principal component is appropriate because it represents the primary direction of the dataset and the direction itself is the very foundation of the PDDP algorithm. However, the first principal component may not always be a good choice, for example when a dataset is similar to the one in Figure 1. In this case the primary direction of the data points is still indicated by the first principal component (shown as v_1), but obviously another principal component (shown as v_2) splits the dataset much better, therefore this principal component, even though not being the first one but just the secon one, should be selected as the input of the PDDP algorithm.

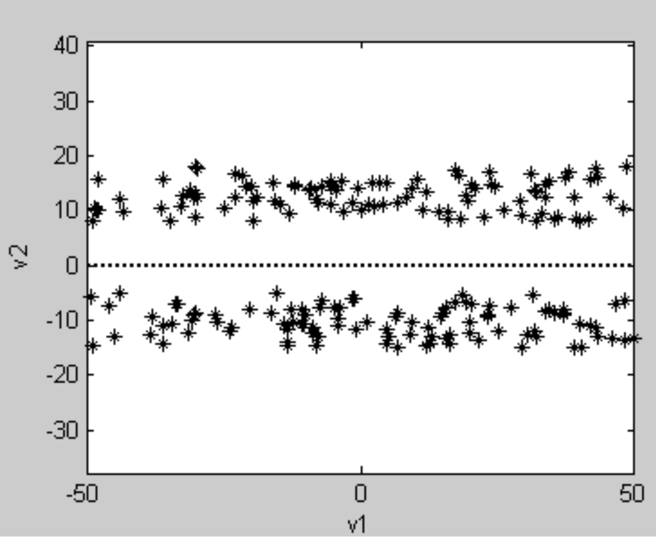

Figure 1 A Special Case of Principal Component Selection

3.3.2 The automatic selection of principal components

The selection of a principal component is easy for supervised PDDP clustering, because we can simply find out from a set of given candidates, for example, the first three principal components, the best one yielding the result closest to the reference. However, when an unsupervised PDDP clustering algorithm is applied, the selection of an appropriate principal component should be done on automatic basis other than manually. In Savaresi et al. (2002) a method that is originally designed for selecting what clusters are to split is deemed to be also helpful for selecting principal components, just after slight modification. The description of the modified algorithm is following.

Suppose the matrices S_0 and V have been worked out from Section 3.2, and a candidate principal component set P = {v_1, v_2, …, v_q} (usually P = { v_1, v_2, v_3}) has been given out.

(1) Write matrix S_0 as $[S_{0,1}, S_{0,2}, \ldots, S_{0,p}]^T$. For each principal component v_j in the given set P, calculate scalar $k_{i,j} = S_{0,i} \cdot v_j$ $(1 \le i \le p, 1 \le j \le q)$. If $k_{i,j} \le 0$, then $k_{i,j} \in K_{j,L}$, otherwise $k_{i,j} \in K_{j,R}$. Write $K_{j,L}$ and $K_{j,R}$ as two row vectors $K_{j,L} = [k_{j,L,1}, k_{j,L,2}, \ldots, k_{j,L,l}]$ and $K_{j,R} = [k_{j,R,1}, k_{j,R,2}, \ldots, k_{j,R,r}]$.

(2) Let $K_{j,L} = K_{j,L} / \min(K_{j,L})$ and $K_{j,R} = K_{j,R} / \max(K_{j,R})$. This normalizes $K_{j,L}$ and $K_{j,R}$ so that all their absolute values range from 0 to 1.

(3) Let scalars $w_{j,L}$ and $w_{j,R}$ be the mean values of $K_{j,L}$ and $K_{j,R}$, respectively, and $w'_{j,L}$ and $w'_{j,R}$ be the mean values of $[(k_{j,L,1}-w_{j,L})^2, (k_{j,L,2}-w_{j,L})^2, \ldots, (k_{j,L,l}-w_{j,L})^2]$ and $[(k_{j,R,1}-w_{j,R})^2, (k_{j,R,2}-w_{j,R})^2, \ldots, (k_{j,R,r}-w_{j,R})^2]$, respectively. Calculate ratio $R_j = \frac{w'_{j,L} + w'_{j,R}}{w_{j,L}^2 + w_{j,R}^2}$.

(4) Select the principal component with the minimum ratio R.

3.4 K-means and Bisect K-means

K-means (MacQueen, 1967) (Pang-ning Tan et al., 2005) is a popular iterative clustering method. The clustering is based on some randomly selected "center points". The number of random points – thus the numebr of obtained clusters - is predefined and determines the number of clusters that the algorithm will output. The basic principle of K-means is as follows:

(1) Randomly select k points (c_1, c_2, …, c_k) from a dataset $S=[S_1,S_2,\ldots,S_p]^T$ in which S_i ($1\le i\le p$) denotes the ith sample. These k random points are viewed as the initial "center points" of k clusters and refined later.

(2) For each sample S_i ($1\le i\le p$), find out a number m, so that for any $j\neq m$ ($1\le m$, $j\le k$), $||S_i-c_m|| \le ||S_i-c_j||$, then $S_i \in C_m$, where $||S_i-c_m||$ and $||S_i-c_j||$ are respectively the distances, for example the Euclidean distances, from S_i to c_m and c_j, and C_m denotes the mth cluster.

(3) Calculate the new center points i.e. the mean values w_1, w_2, …, w_k for the clusters C_1, C_2, …, C_k.

(4) If for each cluster j ($1\le j\le k$), $c_j=w_j$, then stop; otherwise let $c_j=w_j$ for each j, and go to step (2).

K-means algorithm is iteratively convergent, and, if the initial "center points" are selected well, that is to say, they are close to the true center points, then K-means will converge more rapidly, and the clustering result will be more accurate. However, it may not be easy to select good initial center points if one does not know in advance what the distribution of the data points is. This is the reason why to take random points as the initial centers. On the other hand, to apply K-means, the total number of clusters must be determined prior to the clustering.

One kind of K-means, which can be repeatedly applied to form multiple clusters by separating one cluster at a time to get two sub clusters, is called bisect K-means. Similarly, bisect K-means algorithm has the following steps:

(1) Randomly select two “center points”, c_1 and c_2, from the dataset $S=[S_1,S_2,\ldots,S_p]^T$.

(2) If $||S_i-c_1||\leq||S_i-c_2||$, then $S_i \in C_1$; otherwise $S_i \in C_2$, $(1\leq i\leq p)$, where $||S_i-c_1||$ and $||S_i-c_2||$ are the distances, for example the Euclidean distances, from S_i to c_1 and c_2, respectively, and C_1 and C_2 denote the two sub clusters.

(3) Calculates the new center points w_1 and w_2 for the two sub clusters C_1 and C_2.

(4) If $c_1=w_1$ and $c_2=w_2$, then stop; otherwise let $c_1=w_1$ and $c_2=w_2$, and go to step (2).

To get more sub clusters, one can select a cluster, replace dataset S with it, and simply repeat the above steps. Such a procedure can be repeated until a desired number of clusters is obtained.

3.5 Combining PDDP with Bisect K-means

K-means algorithm performs well when the distance information between data points is important to the clustering. However, K-means has an intrinsic disadvantage. The clustering result depends greatly on the selection of initial “center points”. Pang-ning Tan et al. (2005) show different results by applying K-means on the same dataset with different choices of initial “center points”. PDDP has its own weakness, too. Since the partition of PDDP is only on the basis of the projection from the data points to a selected principal direction, the distance information between these data points is ignored.

In spite of the fact that in many cases neither PDDP nor K-means alone is good enough for deriving desirable clustering results, according to the theory of Savaresi and Boley (2001), Savaresi et al. (2002), Savaresi and Boley (2004), the combination of PDDP and bisect K-means keeps the merits of both algorithms, and usually performs better than either single one does. PDDP, although is weak at taking advantage of distance information, can provide bisect K-means good initial center points that are close to true ones, therefore the accuracy of bisect K-means clustering can be improved. The difference between the combined method and the traditional bisect K-means lies in the selection of the initial center points, c_1 and c_2. With the combined method, the two center points of bisect k-means are not selected randomly but according to the clustering result of PDDP, that is to say, c_1 and c_2 should be the sample mean values of the PDDP clusters S_L and S_R, respectively. The combination of PDDP and bisect K-means makes the selection of c_1 and c_2 more reasonable by reducing the risk caused by a random selection.

Figure 2 is a 2-D illustration of the PDDP plus bisect K-means algorithm. In the figure, suppose a 2-D dataset is clustered using the combined method, the data points are represented as blue dots, and their origin is the green dot with coordinates (0, 0). First, by PCA analysis, the origin is moved to the centroid of the dataset (shown as a red dot) along the direction indicated by the dashed arrow, and a principal component is selected with its direction indicated by the black arrow which passes through the new origin and two orange dots. Then, by PDDP, the dataset is separated by another black arrow which passes through the new origin and is perpendicular to the principal direction. The two black arrows actually compose the two coordinates of the new coordinates system. Finally, after PDDP, the centroids of both clusters (shown as two orange dots) are selected as the initial center points of bisect K-means, and the dataset is clustered based on this selection.

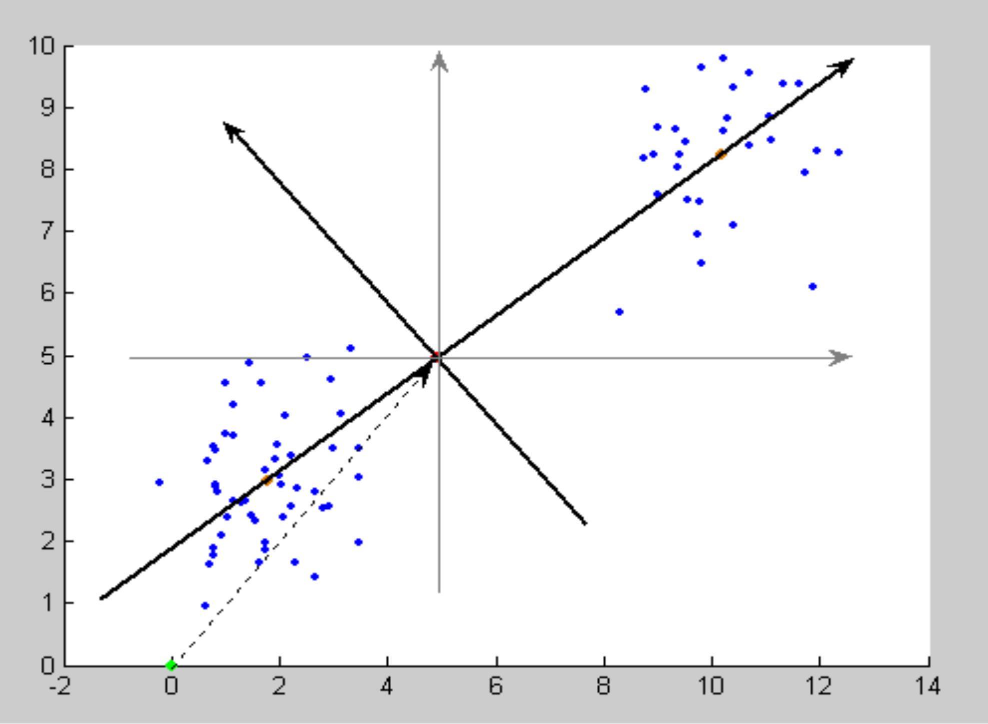

Figure 2 **A 2-D Illustration of PDDP + Bisect K-means**

3.6 The Extraction of Significant Attributes

As already mentioned, the extraction of significant features strongly related to clustering is also a key issue, besides the clustering itself. To achieve this, one should first know the degree of significance of each attribute. Fortunately, principal component analysis itself can also provide quantitative information to measure the significance. A method of extracting the most significant attributes based on PCA analysis is following:

(1) Suppose vector $v_j=[v_{1j},v_{2j},\ldots,v_{Nj}]^T$ is the j-th principal component of S_0 (i.e. column j of V where $S_0=U\sum V$) and v_j is selected to do PDDP. Sort vector v_j in a descending order of $|v_{ij}|$ ($1\le i\le N$) and write it as $v'_j=[v'_{1j},v'_{2j},\ldots,v'_{Nj}]^T$. Since the

significance of each attribute is reflected by the absolute value of the corresponding element in the principal component, now v'_{1j} is the significance coefficient of the most important attribute,v'_{2j} is that of the second most attribute, and so on.

(2) Redo the PDDP + bisect K-means clustering using the reduced principal component $\mathbf{u}_\mathbf{m}$, ($1 \leq m \leq N$), and find out the minimum value of m that outputs the best clustering result that is the closest to a reference result, then the m corresponding attributes are the solution.

3.7 Supervised and Unsupervised Clustering

With a supervised clustering approach, some a priori knowledge such as a pre-defined reference result and the number of clusters can be used to guide the process of clustering. However, such a priori knowledge is not always available before clustering; they may be known only when the clustering is successfully completed. In this case, an unsupervised alternative can be considered when applicable. The PDDP + bisect K-means algorithm is capable of dividing data points into two clusters in either supervised or unsupervised way, as described in the following procedures:

3.7.1 Procedure PCA

```
Procedure PCA
    Input: $p \times N$ data matrix S.
    Output: sorted principal component vector v and index vector x.
    Begin
        Calculate the centralized matrix $S_0$ of S;
        Do singular value decomposition with $S_0$ and get the principal components;
        Select a principal component manually or automatically;
    Sort its elements in the descending order of their absolute values, and get the index of each attribute corresponding to the order;
    Return v (the sorted principal component vector) and x (the index vector);
End
```

3.7.2 Procedure PDDP_Bisect_K-means_Unsupervised

```
Procedure PDDP_Bisect_K-means_Unsupervised
  Input: matrix S, vector v (output of procedure PCA), and vector x (output of procedure PCA).
  Output: two clusters $S_L$ and $S_R$ and the significant attribute set A
  Begin
    Use matrix S and vector v to do PDDP + Bisect K-means clustering, and get two clusters $S_L$ and $S_R$;
    For (i <- 1 to N-1)
        $v_i$ <- v;
        Set the last N-i elements in $v_i$ to 0;
        Use S and $v_i$ to do PDDP + Bisect K-means, and get two clusters $S_{Li}$ and $S_{Ri}$;
        If (($S_{Li}$=$S_L$) and ($S_{Ri}$=$S_R$))
            Break;
        End If
    End For
    A <- the first i indices in x;
```

```
        Return $S_L$, $S_R$, and A;
    End
```

3.7.3 Procedure PDDP_Bisect_K-means_Supervised

Procedure PDDP_Bisect_K-means_Supervised

Input: matrix S, vector v (output of procedure PCA), vector x (output of procedure PCA), and vector c as the reference result of clustering.

Output: two clusters S_L and S_R and the significant attribute set A.

```
Begin
    Get two clusters $S_{Lc}$ and $S_{Rc}$ from matrix S and reference result c;
    err <- p;
    m <- 0;
    For (i <- 1 to N)
        $v_i$ <- v;
        Set the last N-i elements in $v_i$ to 0;
        Use S and $v_i$ to do PDDP + Bisect K-means, get two clusters $S_{Li}$ and $S_{Ri}$ and the clustering result $c_i$;
        Calculate $err_i$, the number of differences between c and $c_i$;
        If ($err_i$ < err)
            err <- $err_i$;
            m <- i;
        End If
    End For
```

```
        $S_L$ <- $S_{Lm}$;
        $S_R$ <- $S_{Rm}$;
        A <- the first m indices in x;
        Return $S_L$, $S_R$, and A;
End
```

4. Experimental Case: Data and Results

This Section is focused on some experimental results about the clustering of the leukemia gene expression dataset mentioned previously. The original dataset S consists of 72 samples (24 ALL, 20 MLL and 28 AML patients distributed in a training dataset of 57 samples and a testing dataset of 15 samples) and each sample is represented by 12,582 gene expression values. The samples are numbered as: #1 - # 20 (ALL in training), #21 - #37 (MLL in training), #38 – #57 (AML in training), #58- #61 (ALL in testing), #62 - #64 (MLL in testing), and #65 - #72 (AML in testing). Dataset S is stored as a 72×12,583 matrix, because there is an extra column, column 12,583, which represents the clustering result presented in Scott et al. (2002). In this column, classes ALL, MLL, and AML are represented as 0, 1, and 2, respectively. This column serves as the reference result of all the following experiments. In other words, the experiment results are compared with the reference, and any different clustering cases are reported as "errors" and analyzed later. As already said, before any experiments, a threshold th = 400 is applied to remove those genes with standard deviation values less than 400, since they are with little possibility to be significant attributes. To verify the effectiveness of the threshold, every experiment is then repeated with th = 0 i.e. all the genes included. The exactly same results and much less execution time show that the threshold applied is reasonable and effective at least in this experimental case. All the experiments are based on the MATLAB implementation of the algorithms described in Section 3.

4.1 The Unsupervised Clustering of Dataset S

With threshold th = 400, the input dataset S becomes a 72×6,611 matrix. With the first principal component and all the 6,611 genes, a clustering result is shown in Table

1, where two initial clusters, S_L and S_R, are obtained. It should be mentioned that, this initial clustering successfully separates ALL and AML with only an exception at sample #3, if we claim that all ALL samples belong to S_L and all AML belong to S_R. This implies that we correctly identify 23 out of 24 ALL and all the 28 out of 28 AML samples, like in Garatti et al. (2007).

	Patient Numbers																			
S_L	1	2	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
S_R	3	22	24	26	27	29	31	38	39	40	41	42	43	44	45	46	47	48	49	50
	Patient Numbers																			
S_L	23	25	28	30	32	33	34	35	36	37	58	59	60	61	62	63	64			
S_R	51	52	53	54	55	56	57	65	66	67	68	69	70	71	72					

Table 1 Unsupervised Clustering Result of Dataset S

The minimum gene set that produces the above result consists of only two genes: #28 (in the original 12,582-attribute dataset) whose name is AFFX-UMGAPDH/M33197_5_at

and

#12,430 with the name 256_s_at.

Table 2 gives out the significance coefficient information about these two genes. The significance coefficients are obtained by taking the absolute values of the corresponding elements in the first principal component, the average coefficient is the mean of the absolute values of all the 6,611 coefficients, and the normalized coefficients, which are used as the contribution indicator of the genes to the clustering, are the quotients of the significance coefficients to the average coefficient.

Gene #	Gene Name	Significance Coeffcient	Average Coefficient	Normalized Coefficient
28	AFFX-HUMGAPDH/M33197_5_at	0.1113	0.0073	15.2466
12	256_s_at	0.0984		13.4795

Table 2 Significant Genes for the Clustering of Dataset S

From Figures 3 and 4, the plotting of the 72 expression values of these two genes, we can visually separate S_L (with relatively low expression values) and S_R (with relatively high expression values) to a certain extent, although a few exceptional cases exist. The rationale of the extraction of these two genes is thus illustrated in such a manner.

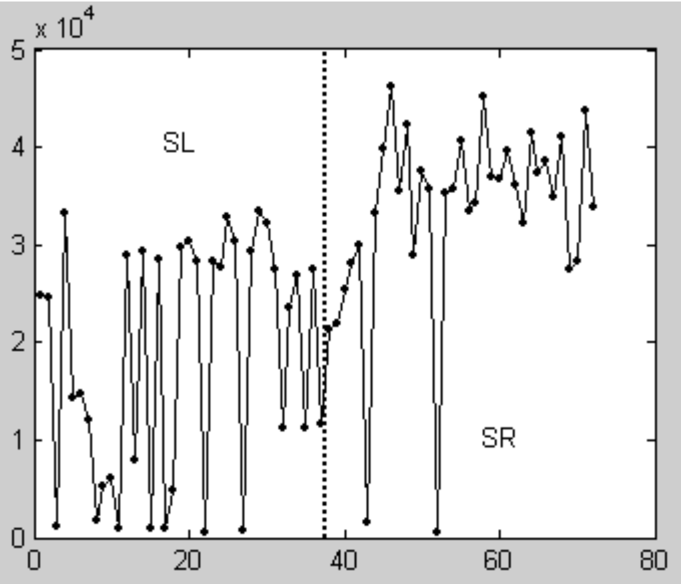

Figure 3 Expression Values of Gene #28

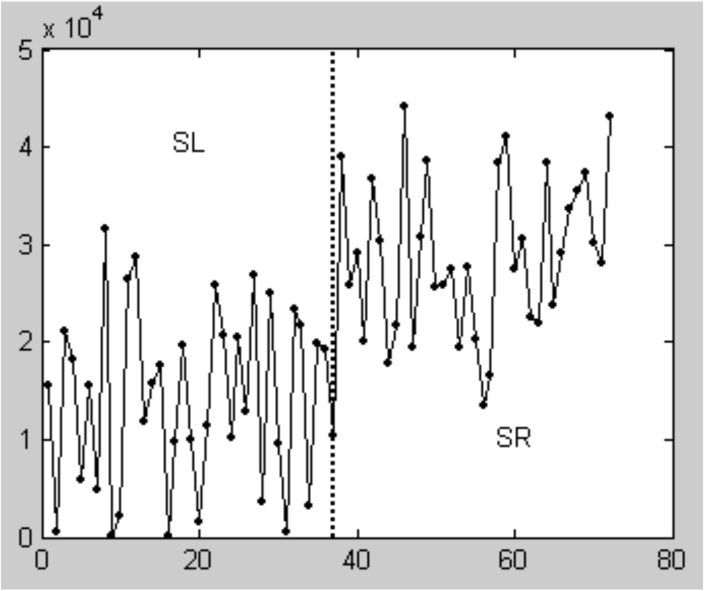

Figure 4 Expression Values of Gene #12,430

It is natural that the initial clustering does not give out any useful information about the MLL samples, because the PDDP based approach only produces two clusters after a single application. For this reason, further clustering is needed to hopefully reveal the aspect of the MLL part.

4.2 The Unsupervised Clustering of Sub Dataset S_L

According to the result of the initial clustering, 37 samples are classified as S_L; among them 23 are actually ALL samples and 14 are MLL.). Clustering of subclass S_L is continued in order to see if the PDDP based approach can successfully identify these ALL samples from the non ALL ones (i.e., according to the reference, the MLL ones, at the first bipartition clustered with ALL, thus closest to such ones than to AML). With the first principal component, 5,962 genes (threshold th = 400), and two significant genes, a result is obtained exactly reproducing the reference, as shown in Table 3, listing the patient numbers and the subclasses that they belong to. Based on Table 3, we claim that S_{LL} and

S_{LR} are actually ALL and a part of MLL, respectively. Table 4 gives out the two significant genes and quantifies their contribution to the clustering. Figures 5 and 6 plot the 37 expression values of these two genes.

	Patient Numbers																			
SLL	1	2	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	58
SLR	21	23	25	28	30	32	33	34	35	36	37	62	63	64						
	Patient Numbers																			
SLL	59	60	61																	
SLR																				

Table 3 Unsupervised Clustering Result of Sub Dataset S_L

Gene #	Gene Name	Significance Coefficient	Average Coefficient	Normalized Coefficient
7,	33412_at	0.1533	0.0072	21.2917
11,924	769_s_at	0.1083		15.0472

Table 4 Significant Genes for the Clustering of Sub Dataset S_L

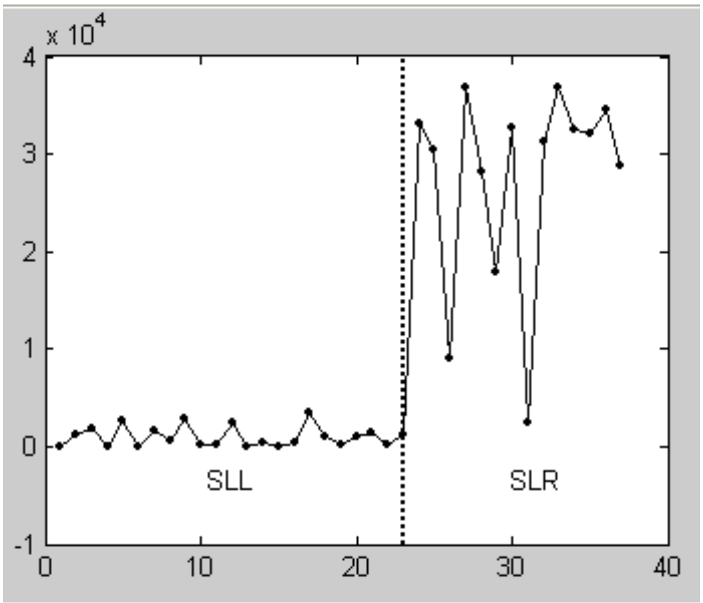

Figure 5 Expression Values of Gene #7,754

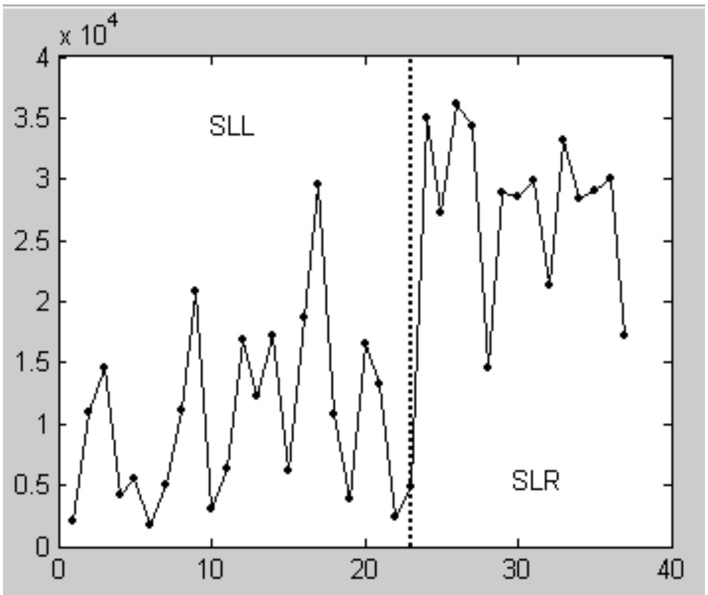

Figure 6 Expression Values of Gene #11,924

4.3 The Unsupervised Clustering of Sub Dataset S_R

Since the initial clustering of dataset S is insufficient for identifying the MLL samples, a similar clustering of the subclass S_R is then performed to see whether those MLL samples can be separated successfully. According to the result of the initial clustering, 35 samples are classified as S_R. Among them are 28 AML, 6 MLL, and one misclassified ALL. With the first principal component and 6,191 genes (threshold th = 400), the result is shown in Table 4.5. The minimum gene set with the clustering result in this table consists of 219 genes; they are not reported in this paper.

The clustering seems unsuccessful, with many AML samples and all the MLL samples clustered together into S_{RL}. However, an interesting observation is that no MLL sample is clustered into S_{RR} as shown in Table 5 with MLL patients shaded in grey.

	Patient Numbers																	
SRL	3	22	24	26	27	29	31	42	43	47	48	50	52	55	68	70	72	
SRR	38	39	40	41	44	45	46	49	51	53	54	56	57	65	66	67	69	71

Table 5 The Unsupervised Clustering Result of Sub Dataset S_R

4.4 The Supervised Clustering of Sub Dataset S_{RL}

Because all the 6 MLL samples are classified as S_{RL} in Section 4.3, it may be interesting to continue clustering the sub cluster S_{RL}. With the first principal component and 5,877 genes (threshold th = 400), an unsupervised result with two errors is obtained. The minimum gene set for this result consists of 103 genes which are not reported in this paper. However, when the clustering is performed under the supervision of the reference result, a better result is obtained with only one error at patient #3, as shown in Table 6, listing the patient numbers and their clusters according to this supervised clustering. The minimum gene set for this result consists of 9 genes. They are listed in Table 7.

	Patient Numbers									
S RLL	3	22	24	26	27	29	31			
S RLR	42	43	47	48	50	52	55	68	70	72

Table 6 Supervised Clustering Result of Sub Dataset S_{RL}

Gene #	Gene Name	Significance Coefficient	Average Coefficient	Normalized Coefficient
12	319_g_at	0.1106	0.0080	13.8250
31	AFFX-HSAC07/X00351_5_at	0.1106		13.8250
32	AFFX-HSAC07/X00351_M_at	0.0995		12.4375
7,	33412_at	0.0993		12.4125
1,	33516_at	0.0989		12.3625
28	AFFX-HUMGAPDH/M33197_5_at	0.0985		12.3125
1,	35083_at	0.0950		11.8750
8,	36122_at	0.0940		11.7500
3, 634	39318_at	0.0933		11.6625

Table 7 Significant Genes for the Clustering of Sub Dataset S_{RL}

5. Discussion

5.1 Discussion about the Experimental Results

5.1.1 Discussion about the clustering results

According to the clustering results in Section 4, the leukemia dataset S can be clustered as the following hierarchy:

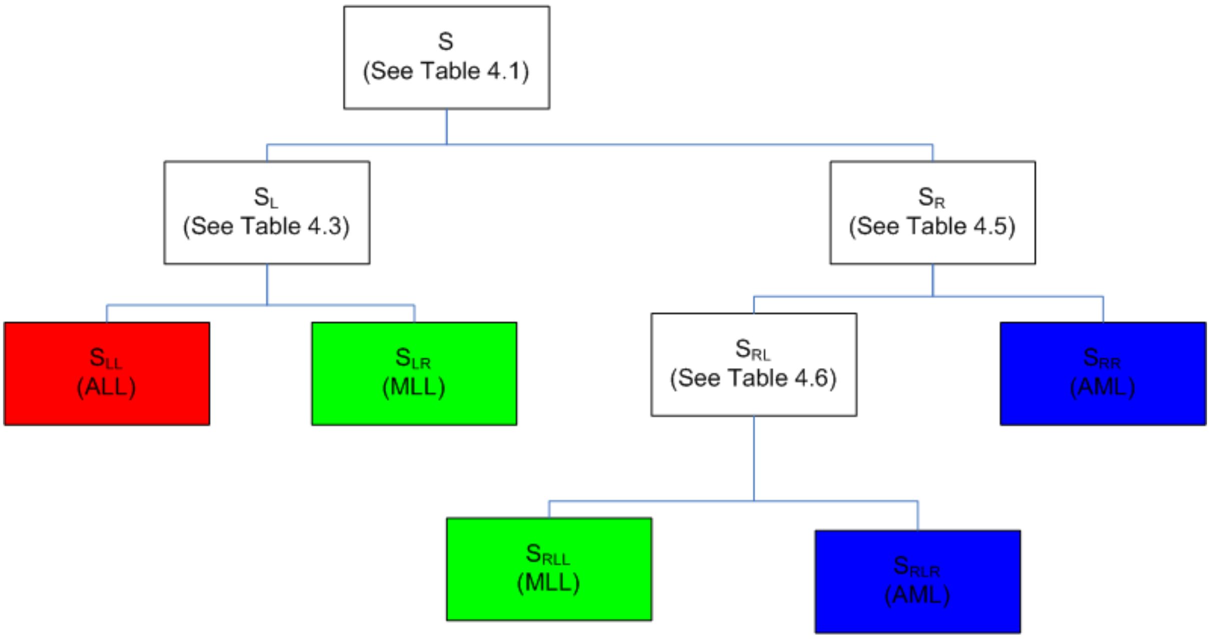

Figure 7 **The Hierarchy of the Leukemia Dataset** (referred Tables inherit prefix 4. in order to remind they are in Results Chapter 4)

In Figure 7, if we name cluster S_{LL} as ALL, clusters S_{LR} and S_{RLL} together as MLL, and clusters S_{RLR} and S_{RR} together as AML, then there is only one error occuring in the whole set of patient with such supervised aggregation. From Table 3, almost all the 24 ALL patients are identified in cluster S_{LL}, except patient #3 eventually misclassified into cluster S_{RLL}; this is the only error that occurs. It may be due to impreciosion of the algorthm, but also to orginal misclassificaton of the data, or just be a borderline subjet difficult to classify being closed to another class at least in the reduced used subspace (Garatti et al, 2007).

From tables 3 and 6, 14 MLL patients are identified in cluster S_{LR} and other 6 are identified in S_{RLL}; these two clusters include all the MLL patients without any misclassification.

From tables 5 and 6, 18 AML patients are identified in cluster S_{RR} and the other 10 in cluster S_{RLR}; these two clusters include all the AML patients without any misclassification.

Interestingly enaough, except for ALL, both MLL and AML patients are divided into two sub clusters. This implies that there might exist other sub types for MLL and AML. In fact, on the very same subset, Golub et al. (1999) labeled only two sub types of leukemia (ALL and AML) while Scott et al. (2002) detailed proposing the three sub types (ALL, MLL, and AML) analyzed in this paper.

5.1.2 Discussion about the significant genes

First, by reviewing the gene extraction results in Section 4, we see that different levels of expression values of just the 2 genes #28 (AFFX-HUMGAPDH/M33197_5_at) and #12,430 (256_s_at) are already enough to well separate ALL and AML patients. Second, in the initial clustering of the dataset, most MLL data are shown closer to ALL than AML, implying that MLL and ALL share similarity to a great extent: in fact, they were classified together in the same class by Golub et al. (1999). The difference between ALL and MLL is then very well revealed by just 2 more genes #7,754 (33412_at) and #11,924 (769_s_at). On the other hand, a small portion of MLL data are shown closer to AML, showing that some MLL and AML cases may have common characteristics. The size of the minimum set of genes which separate MLL from AML is very large, implying that genetically diagnosing AML-like MLL patients may be more difficult than that of ALL-like MLL patients. Finally, the contribution of genes to the corresponding clustering results is quantified so that the significance of them can be compared quantitatively. For examples, gene #28 (normalized significance coefficient (NSC) = 15.2466) and #12,430 (NSC = 13.4795) are almost equally significant to the discrimination between ALL and AML, while gene #7,754 (NSC = 21.2917) appears to be more significant than #11,924 (NSC = 15.0472) to the discrimination between MLL and ALL, and so on.

5.2 Conclusion

With the combined approach of PDDP and bisect K-means, 72 leukemia patients are successfully clustered as ALL, MLL and AML, respectively. Among all the 12,582 genes, the most discriminating ones that are responsible for the clustering are efficiently discovered. Furthermore, both the clustering of patients and the discovering of significant genes are performed automatically to a great extent, and depend merely on the gene expression data which can be obtained conveniently by using the popular DNA micro array technology.

In conclusion, the combination of PDDP and bisect K-means is an efficient approach for the clustering of the leukemia patient dataset described in this paper, and hopefully also efficient for other similar datasets. Moreover, the significant genes discovered among tens of thousands of genes may provide very important information for the diagnosis of leukemia. The same approach reveals to be useful to other tumor classifications, like pancreatic ones (in preparation), even if not all case: it does not work for instance in discriminating breast cancer. This is understandable by considerign that the proposed approach works in a quasi-linear partitioning, that is not in general appropriate to any data set. When woring, like in this paper, it offers a powerful simlple approach to gain immediateknowledge about the few genes mainly involved in classification, thus possibly offering hints in understandig pathophysiology and suggesting and monitoring teraphy, beyond the scope of this very paper

6. Acknowledgements

To colleagues , postdocs and students involved in the project at various title

7. References

[1] Golub, T.R., D.K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J.P. Mesirov, H. Coller, M. Loh, J.R. Downing, M.A. Caligiuri, C.D. Bloomfield and E.S. Lander: "Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring". Science, 286:531-537, October 1999.

[2] S Garatti, S Bittanti, D Liberati, A Maffezzoli An unsupervised clustering approach for leukaemia classification based on DNA micro-arrays data, Intelligent Data Analysis 11 (2), 175-188, 2007.

[3] van't Veer LJ, Dai HY, van de Vijver MJ, He YDD, Hart AAM, Mao M, Peterse HL, van der Kooy K, Marton MJ, Witteveen AT, Schreiber GJ, Kerkhoven RM, Roberts C, Linsley PS, Bernards R and Friend SH: "Gene Expression Profiling Predicts Clinical Outcome of Breast Cancer". Letters to Nature, Nature, 415:530-536, 2002.

[4] Pomeroy SL, Tamayo P, Gaasenbeek M, Sturla LM, Angelo M, McLaughlin ME, Kim JY, Goumnerova LC, Black PM, Lau C, Allen JC, Zagzag D, Olson JM, Curran T, Wetmore C, Biegel JA, Poggio T, Mukherjee S, Rifkin R, Califano A, Stolovitzky G, Louis DN, Mesirov JP, Lander ES and Golub TR: "Prediction of Central Nervous System Embryonal Tumour Outcome Based on Gene Expression". Letters to Nature, Nature, 415:436-442, January 2002.

[5] Alon U, Barkai N, Notterman DA, Gish K, Ybarra S, Mack D and Levine AJ: "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, 96:6745-6750, 1999.

[6] De Cecco L, Marchionni L, Gariboldi M, Reid JF, Lagonigro MS, Caramuta S, Ferrario C, Bussani E, Mezzanzanica D, Turatti F, Delia D, Daidone MG, Oggionni M, Bertuletti N, Ditto A, Raspagliesi F, Pilotti S, Pierotti MA, Canevari S, and Schneider C: "Gene expression profiling of advanced ovarian cancer: characterization of a molecular signature involving fibroblast growth factor 2 ". Oncogene, 23(49):8171-8183, October, 2004.

[7] Dinesh Singh, Phillip G. Febbo, Kenneth Ross, Donald G. Jackson, Judith Manola, Christine Ladd, Pablo Tamayo, Andrew A. Renshaw, Anthony V. D'Amico, Jerome P. Richie, Eric S. Lander, Massimo Loda, Philip W. Kantoff, Todd R. Golub and William R. Sellers: "Gene Expression Correlates of Clinical Prostate Cancer Behavior". Cancer Cell, 1:203-209, March, 2002.

[8] Eng-Juh Yeoh, Mary E. Ross, Sheila A. Shurtleff, W. Kent Williams, Divyen Patel, Rami Mahfouz, Fred G. Behm, Susana C. Raimondi, Mary V. Relling, Anami Patel, Cheng Cheng, Dario Campana, Dawn Wilkins, Xiaodong Zhou, Jinyan Li, Huiqing Liu, Ching-Hon Pui, William E. Evans, Clayton Naeve, Limsoon Wong and James R. Downing: "Classification, Subtype Discovery, and Prediction of Outcome in Pediatric Acute Lymphoblastic Leukemia by Gene Expression Profiling". Cancer Cell, 1:133-143, March, 2002.

[9] Gavin J. Gordon, Roderick V. Jensen, Li-Li Hsiao, Steven R. Gullans, Joshua E. Blumenstock, Sridhar Ramaswamy, William G. Richards, David J. Sugarbaker and Raphael Bueno: "Translation of Microarray Data into Clinically Relevant Cancer Diagnostic Tests Using Gege Expression Ratios in Lung Cancer And Mesothelioma". Cancer Research, 62:4963-4967, 2002.

[10] Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A, Boldrick JC, Sabet H, Tran T, Yu X, Powell JI, Yang L, Marti GE, Moore T, Hudson J Jr, Lu L, Lewis DB, Tibshirani R, Sherlock G, Chan WC, Greiner TC, Weisenburger DD, Armitage JO, Warnke R, Levy R, Wilson W, Grever MR, Byrd JC, Botstein D, Brown PO and Staudt LM: "Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling". Nature, 403:503-511, February 2000.

[11] Scott A. Armstrong, Jane E. Staunton, Lewis B. Silverman, Rob Pieters, Monique L. den Boer, Mark D. Minden, Stephen E. Sallan, Eric S. Lander, Todd R. Golub and Stanley J. Korsmeyer: "MLL Translocations Specify A Distinct Gene Expression Profile that Distinguishes A Unique Leukemia". Nature Genetics, 30:41-47, January 2002.

[12] Hand, D., H. Mannila, P. Smyth (2001): “Principles of Data-Mining”. The MIT press, Cambridge, Massachusetts, USA.

[13] O'Connel M.J.: “Search Program for Significant Variables”. Computer Physics Communications, 1974. 8: p. 49-55.

[14] Wall ME, Rechtsteiner A and Rocha LM: “Singular value decomposition and principal component analysis". A Practical Approach to Microarray Data Analysis. (Berrar DP, Dubitzky W, Granzow M, eds.), pp. 91-109, Kluwer:Norwell, MA (2003).

[15] Boley, D.L. (1998): "Principal Direction Divisive Partitioning". Data Mining and Knowledge Discovery, 2(4), 325-344.

[16] Savaresi, S., Boley, D., Bittanti, S. and Gazzaniga, G. (2002): “Choosing the cluster to split in bisecting divisive clustering algorithms”. Second SIAM International Conference on Data Mining (SDM'2002)

[17] Canasai Kruengkrai, Virach Sornlertlamvanich and Hitoshi Isahara: “Refining A Divisive Partitioning Algorithm for Unsupervised Clustering”. The 3rd International Conference on Hybrid Intelligent Systems (HIS'03), December 14-17, 2003

[18] Pang-ning Tan, Michael Steinbach and Vipin Kumar: “Introduction to Data Mining”, Addison Wesley Publishing Company, 2005.

[19] J. MacQueen: “Some methods for classification and analysis of multivariate observations”. L. M. LeCam and J. Neyman, editors, Proceedings Fifth Berkeley Symposium on Math. Stat. and Prob., pages 281--297. University of California Press, 1967.

[20] Savaresi, S.M. and D.L. Boley (2001): "On the performance of bisecting K-means and PDDP". 1st SIAM Conference on Data Mining, Chicago, IL, USA, paper n.5, pp.1-14.

[21] Savaresi, S.M., D.L. Boley, S. Bittanti and G. Gazzaniga (2002): "Cluster selection in divisive clustering algorithms". 2nd SIAM International Conference on Data Mining, Arlington, VI, USA, pp.299-314.

[22] Savaresi, S.M. and D.L. Boley (2004): “A Comparative Analysis on the Bisecting K-Means and the PDDP Clustering Algorithms”. International Journal on Intelligent Data Analysis, 8(4), pp. 345-362.