A Comparative Investigation of *K*-means and Partition Around Medoid Methods of Clustering - a Case Study with Acute Lymphoblastic Leukemia Data

Khaled I. A. Almaghri[*] and S. Chakraborty

Department of Statistics, Dibrugarh University

Dibrugarh -786004, *Assam, India.*

**corresponding Author*

دراسة مقارنة بين طرق التحليل العنقودي المجمعة حول الوسط وحول الوسيط باستخدام بيانات سرطان الدم الليمفاوي الحاد كحالة دراسية

مجلة جامعة فلسطين للأبحاث والدراسات – المجلد السادس العدد الثاني
يونيو 2016

ABSTRACT

Clustering methods are important tool in data mining. The main challenge of clustering is to select the suitable method to be used for a given data set and the estimation of the number of clusters in the data set, especially in case of the unsupervised data. In this paper, a comparison between two important partitioning clustering methods namely the *K*-means and the Partition Around Medoid (PAM) have been considered and a special index for each has been used to estimate number of clusters. Also different indices of internal validation and stability measures have been used to compare these two methods to evaluate their performance by using these indices.

Internal validation and stability measures have been used to compare between *K*-means and PAM for B-cells and T-cells and it has been found that for B-cells the K-means performs better than PAM by Connectivity, Dunn, Silhouette, APN, ADM, FOM indexes and PAM perform better than K-means by AD index. For T-cells, PAM performs better than K-means by Connectivity index and *K*-means performs better than PAM by Dunn, Silhouette, APN, AD, ADM, FOM indices.

Keywords: B-cells, T-cells, *K*-means, PAM, Calinski, Silhouette, Connectivity, Dunn, APN, AD, ADM, FOM.

الملخص باللغة العربية

طرق التحليل العنقودي مهمة جداً لفهم البيانات , يعتبر أهم تحدي لاجراء التحليل العنقودي هو اختيار الطريقة المناسبة للتحليل العنقودي وكذلك تقدير عدد العناقيد في البيانات خاصة في البيانات الغير مصنفة أصلاً, وفي هذا البحث قام الباحثان بالمقارنة بين أحد أهم الطرق لتقسيم البيانات وهما طريقة التحليل العنقودي المجمعة حول الوسط وطريقة التحليل العنقودي المجمعة حول الوسيط, وتم استخدام مؤشرات

قياسية لكل طريقة للتنبؤ بعدد العناقيد في كل طريقة وأيضاً تم استخدام عدة مؤشرات لقياس التحقق الداخلي والثبات في كل طريقة لتقييم النتائج في كل منهما وكذلك لاختبار كفاءة كل مؤشر من المؤشرات.
كما تم فحص الاستقرار الداخلي لكل طريقة ومن خلال هذا الفحص تبين أن طريقة التحليل العنقودي المجمعة باستخدام المعايير حول الوسط كانت أكفأ
)Connectivity, Dunn, Silhouette, APN, ADM, FOM ،(
عند التطبيق على)AD بينما كانت طريقة التحليل العنقودي المجمعة حول الوسيط أكفأ باستخدام معيار (خلايا بي.
بينما طريقة التحليل العنقودي المجمعة حول الوسط كانت أكفأباستخدام المعايير (Dunn, Silhouette, APN, AD, ADM, FOM(بينما طريقة التحليل العنقودي المجمعة حول الوسيط كانت أكفأ باستخدام معيار(Connectivity(عند التطبيق على خلايا تي.

1. Introduction

K-means and Partition Around Medoid (PAM) and especial indices for each has been presented, both methods are partitioning methods and they attempt to minimize the distance between objects inside a cluster and these objects inside the same cluster should be similar while dissimilar objects are placed in different clusters.

The main objective of the present study is to compare two nonhierarchical clustering methods, *K*-means and Partition Around Medoid by using the Calinski index for the former and Silhouette width for the later and carrying out internal and stability validation for both. Acute Lymphoblastic Leukemia data with B-cells and T-cells subsets of Ritz Laboratory (Sabina et al., 2004 [21]) have been considered for implementing the objective.

2. Material and Methods

Acute Lymphoblastic Leukemia data set taken from Ritz Laboratory (Sabina et al., 2004 [21]) consists of micro arrays from 128 different individuals with acute lymphoblastic leukemia (ALL). The data available in R data base have already been normalized using Robust Multichip Average (rma) (R manual documentation, 2012 [20], Irizarry et al., 2003 [11]).

The two nonhierarchical clustering methods have been implemented using the R software package with Manhattan distance as the data set comprises both continuous as well as categorical data.

2.1 Data Set: Acute Lymphoblastic Leukemia (ALL) Data

This data frame contains observations on: (i) Patient IDs, (ii) Date of diagnosis, (iii) Sex of the patient (sex), (iv) Age of the patient in years (age), (v) type and stage of the disease: 'B' indicates B-cell ALL while 'T' indicates T-cell ALL (BT), (vi) ‘Remission’: a

factor with two levels, either 'CR' indicates that remission was achieved or 'REF' indicating that the patient was refractory , and remission was not achieved (remission), (vii) 'CR': a vector with the following values: 1: "CR", remission; achieved; 2: "DEATH IN CR", patient died while in remission; 3: "DEATH IN INDUCTION", patient died while in induction therapy; 4: "REF", patient was refractory to therapy (CR), (viii) the date on which remission was achieved, (ix) a logical vector indicating whether t (4; 11) translocation was detected (t411), (x) a logical vector indicating whether t (9; 22) translocation was detected (t922), (xi) a vector indicating the various cytogenetic abnormalities that were detected (cyton), (xii) the assigned molecular biology of the cancer (molb), (xiii) Fusion protein for those with BCR\/ABL which of the fusion proteins was detected, 'p190', 'p190\/p210', 'p210' (fusionp), (xiv) the patient's response to multidrug resistance, either 'NEG', or 'POS' (mdr), (xv) 'kinet' ploidy, either diploid or hyperd (kinet), (xvi) a vector indicating whether the patient had neither continuous complete remission nor not (ccr), (xvii) a vector indicating whether the patient had relapse or not (relapse), (xviii) a vector indicating whether the patient receive a bone marrow transplant or not (transplant), and (xix) follow-up data with 10 possible value 1 to 10 (f.u). The possible values of fu are:

1. "AUBMT \/ REL": autologous bone marrow transplant and subsequent relapse,
2. "BMT \/ CCR": allogeneic bone marrow transplant and still in continuous complete remission,
3. "BMT \/ DEATH IN CR": after allogeneic bone marrow transplant patient died without relapsing,
4. "BMT \/ REL": after allogeneic bone marrow transplant patient relapsed,
5. "CCR": patient was in continuous complete remission,
6. "CCR \/ OFF": patient was in continuous complete remission but off-protocol for some reasons,
7. "DEATH IN CR": died when in complete remission,
8. "MUD \/ DEATH IN CR": unrelated allogeneic bone marrow transplant and death without relapsing,
9. "REL": relapse, and
10. "REL \/ SNC": relapse occurred at central nervous system,

The last variable is (xx) a logical vector indicating whether the cytogenesis was normal (citog).

The data have been presented in the form of an 'exprSet' object which is suitable for implementation and comparison in many of clusters algorithms (Kumar and Sharma, 2011 [16]; Jonathan et al., 2010 [13]) because one can extract subsets from this dataset as Acute Lymphoblastic Leukemia caused by different causes like T.cells, B.cells.

The variable BT gives information about the type (B or T) and stages of the disease (five stages for each type). So from the ALL data set two distinct subsets with respect to two covariates namely T cells and B cells have been extracted for independent investigation using the clustering algorithms.

The values of all the variables in the 95^{th} and the 128^{th} rows of the data set are missing. As such effectively the ALL dataset comprises observations of 126 individuals, more over in the present work four variables namely the variables Patient IDs, date of diagnosis, age of the patient in years and date on which remission was achieved have been omitted before the analysis as they are not relevant for the present investigation. Therefore, in the current work, 126 observations (rows) with only 16 out of 20 variables have been considered for the analysis.

2.2 Distance and indices

Manhattan distance which is a non Euclidean distance between two objects $\mathbf{x_i}$ and $\mathbf{x_j}$, r is the number of observations and computed as: (Kaufman and Rousseeuw, 2005 [15])

$$d_{Manh}(\mathbf{x_i}, \mathbf{x_j}) = \left[\sum_{g=1}^{r} \left| x_{ig} - x_{jg} \right| \right] \qquad (1)$$

is the preferred distance measure when data set contain both continuous and categorical data. It is formally known as l_1 norm.

The **cluster validation index of Calinski and Harabasz** (1974 [4]) is defined as

$$Ch(K) = \frac{[traceB / K - 1]}{[traceW / N - K]} \text{ for } K \in N \qquad (2)$$

where B denotes the error sum of squares between different clusters (Inter cluster) and W is the squared difference of all objects in a cluster from their respective cluster center (intra cluster), N is the number of clustered point, K is the number of clusters. The maximal achieved index value indicates the best clustering method for the data (Calinski and Harabasz, 1974 [4]).

2.3 Internal validation and stability measures for clusters

The internal validation measures reflect the compactness, connectness and separation of the cluster partitions. It's very important to know that the internal methods of cluster validation don't provide a definite guide to the number of cluster (pp. 246, Everitt, 2011 [10]).

2.3.1 Internal validation measures

I. Connectivity:

It measures the extent to which observations are placed in the same cluster as their nearest neighbors in the data and can be computed as:

Let $nn_{i(j)}$ as the j^{th} nearest neighbor of observation i.

Let $x_{i,nn_{i(j)}}$ be zero if i and $\mathrm{nn}_{\mathrm{i(j)}}$ are in the same cluster and $1/j$ otherwise.

For a particular clustering partition $C=\{C_1,\cdots,C_k\}$ of the N observations into k disjoint clusters. Then the connectivity is defined as:

$$\mathrm{Conn(C)}=\sum_{\mathrm{i=1}}^{\mathrm{N}}\sum_{\mathrm{j=1}}^{\mathrm{L}}\mathrm{x}_{\mathrm{i,nn_{i(j)}}} \qquad (3)$$

where L is a parameter that determines the number of neighbors that contribute to the connectivity measures.

Interpretation: $\mathrm{Conn(C)}$ *has a value between 0 and* ∞ *and it should be minimized.*

II. Silhouette width:

The silhouette value measures the degree of confidence in the clustering assignment of a particular observation. It is defined as:

$$S(i)=\frac{b_i-a_i}{\max(b_i,a_i)} \qquad (4)$$

where a_i is the average distance between i and all other observations in the same cluster and it can be defined as:

$$a_i=\frac{1}{n(C(i))}\sum_{j\in C(i)}dist(i,j),$$

b_i is the average distance between i and the observations in the nearest neighboring cluster and it can be defined as:

$$b_i=\min_{C_k\in C\setminus C(i)}\sum_{j\in C_k}\frac{dist(i,j)}{n(C_k)}$$

$C(i)$ is the cluster containing i, $dist(i,j)$ is the distance between observations i, j. In the current investigation the suitable distance is the Manhattan distance $n(C)$ which is the cardinality of cluster C. and our data set contain both continuous and categorical data.

Interpretation: Silhouette width lies in the interval [-1, 1] and should be maximized.

The average of $S(i)$ for all objects i in a cluster, which is called the *average silhouette width of that cluster*.

The average of $S(i)$ for $i=1,2,\cdots,n$ is called the average *silhouette width for the entire data set*, and can be used for the selection of the "best" value of k, by choosing that k for which *silhouette width* is highest.

Silhouette Coefficient (SC) is defined as the maximum Silhouette width for entire dataset. The values of (SC) lie between 0 and 1 and are usually interpreted as follows:

- (0.7 - 1.0) A strong structure has been found.
- (0.5 - 0.7) A reasonable structure has been found.

- (0.26 – 0.5) The structure is weak and could be artificial and there is a need to try other additional methods of clustering to such datasets.
- (≤ 0.25) No substantial structure has been found.

(See Kaufman and Rousseeuw, 2005 [15]; Kumar and Sharma, 2009 [16]; Swami and Jain, 2006 [24]; Anja et al., 1997 [1] for details)

III. Dunn index:

The Dunn index is the ratio of the smallest distance between observations not in the same cluster to the largest intra cluster distance and is given by

$$D(C) = \frac{\min\limits_{C_K, C_L \in C, C_K \neq C_L} \left(\min\limits_{i \in C_K, j \in C_L} dis(i, j) \right)}{\max\limits_{C_m \in C} \text{diam}(C_m)} \quad , \qquad (5)$$

where $\text{diam}(C_m)$ is the maximum distance between observations in cluster C_m.

Interpretation: *The Dunn index has a value between* 0 *and* ∞ *it should be maximized*.

The above three indices have been implemented in R using the function *clValid* of *clValid* library.

2.3.2 Stability measures

For a data set having M observations (rows) per variable and N variables (columns) stability measures implemented in the *clValid* library compares the clustering outputs based on the full data with the clustering based on the data with one column removed one at a time (Datta and Datta, 2003 [7]; Yeung et al. 2001 [25]). It has been shown that these measures provide good results if the data are highly correlated. The four measures of stability available in *clValid* library are the average proportion of non-overlap (APN), the average distance (AD), the average distance between means (ADM), and the figure of merit (FOM).

I. Average proportion of non overlap (APN)

The APN is defined as:

$$APN(C) = \frac{1}{MN} \sum_{i=1}^{N} \sum_{i=1}^{M} \left(1 - \frac{n(C^{i,l} \cap C^{i,0})}{n(C^{i,0})} \right) \quad , \qquad (6)$$

where:

M : Number of observations (rows) per variable.

N : Number of Variables (columns)

$C^{i,l}$: Cluster containing observation i where the clustering is based on the dataset with column l removed.

$C^{i,0}$: Cluster containing observation i using the original clustering based on full data

Interpretation: The values of the APN lies in the interval [0, 1], with values close to zero corresponding with highly consistent clustering results.

II. Average distance (AD)

The AD is defined as:

$$AD(C) = \frac{1}{MN}\sum_{i=1}^{N}\sum_{i=1}^{M}\frac{1}{n(C^{i,0})n(C^{i,l})}\left[\sum_{i\in C^{i,0}, j\in C^{i,l}} dis(i,j)\right] \quad (7)$$

Interpretation: AD has a value between 0 to ∞. The smaller values are preferred.

III. Average Distance between Means (ADM)

The ADM is defined as:

$$ADM(C) = \frac{1}{MN}\sum_{i=1}^{N}\sum_{i=1}^{M} dis(\bar{x}_{C^{i,l}}, \bar{x}_{C^{i,0}}) \quad (8)$$

where $\bar{x}_{C^{i,0}}$ is the mean of the observations in the cluster that contains all columns, and $\bar{x}_{C^{i,l}}$ is the mean of the observations in the cluster that contains the data with removed column.

Interpretation: ADM like the AD has a value between 0 to ∞. The smaller values are preferred.

IV. Figure of Merit (FOM)

The FOM is defines as:

$$FOM(l,C) = \sqrt{(\frac{1}{N}\sum_{k=1}^{k}\sum_{i\in C_k(l)} dis(x_{i,l}, \bar{x}(l))} \quad (9)$$

Where $x_{i,l}$ is the value of the i^{th} observation in the l^{th} column. $\bar{x}_{C_k}(l)$ is the average of the cluster $C_k(l)$.

Interpretation: FOM takes value between 0 to ∞. The smaller values are preferred.

2.4. *K*-means and Partition around Medoid (PAM) algorithm

***K*-means** is a popular algorithm because it is used for large scale clustering projects and it accesses the original data. The algorithm seeks to minimize Error Sum of Squares (ESS) and the procedure stops when no further reassignment reduces ESS.

***K*-means algorithm**:

I. Input $l = \{x_i, i = 1,2,\cdots,n\}$, *K*= number of clusters.

II. Do one of the following :

- Start with initial random assignment of the items into *K* clusters and for cluster *m* compute its current centroid as: $\bar{x}_m, m = 1,2,\cdots,K$.
- Pre-specify the squared *K* cluster centroid as $\bar{x}_m, m = 1,2,\cdots,K$

III. Compute the squared Euclidean distance of each item to its current cluster centroid as:

$$\text{ESS} = \sum_{m=1}^{K}\sum_{c(i)=m}(x_i - \bar{x}_m)^T(x_i - \bar{x}_m) \quad (10)$$

Where $\bar{x}_m$ is the m^{th} cluster centroid and $\boldsymbol{c(i)}$ is the cluster containing $\boldsymbol{x_i}$.

IV. Reassign each item to its nearest cluster centroid so that ESS is reduced to magnitude. After each assignment update the clusters centroid.

V. Repeat steps 3 and 4 until no further reassignment of items takes place.

Remark 1. *K*-means algorithm is probably the most widely applied nonhierarchical clustering techniques (Kaufman and Rousseeuw, 2005 [15]; Brito et al., 2007 [3]). It can be implemented easily using an update equation for the centroid coordinates, if object i is moved from cluster v to cluster w, the new centroid are given by:

$$\bar{x}_f(v') = \frac{1}{n_v - 1}(n_v \bar{x}_f(v) - x_{if}) \qquad (11)$$

and

$$\bar{x}_f(w') = \frac{1}{n_w - 1}(n_w \bar{x}_f(w) - x_{if}) \qquad (12)$$

where:

- v' and w' represent the new clusters
- $\bar{x}_f(v'), \bar{x}_f(w')$ are the centroid of cluster v, w respectively.
- n_v, n_w are numbers of objects in clusters v, w respectively.

As the centroid is the point which minimizes the sum of squares of distances, the total sum of squares will decrease by an even larger quantity. (See pp.424, Izenman, 2008 [12]; Dean and Richaed, 2002 [8]; Kumar and Sharma, 2011 [16]; Qin, 1999 [19] for details). One of the main limitations of *K*-means algorithm is the effect of outliers on the results.

Partition around Medoid (PAM) is a modified form of *K*-means and a more robust than *K*-means, PAM depends on the sum of distances between the medoid* and the other cluster members. This sum should be the minimum (*clValid* library R, R manual documentation, 2012 [20]). The main advantages of this method the lies in its computation - and findings which are truly representative of the observations within a given cluster. (Cluster analysis [6])

* Clusters are typically represented by centrotypes which are objects in the cluster having maximum within average similarity (or minimum dissimilarity). Medoid is one such centrotypes which is characterized by having minimum absolute distance among other members of that cluster.

Algorithm of partitioning-around-Medoids clustering (pp.426, Izenman, 2008 [12])

1. Input: Proximity matrix $\mathbf{D} = (d_{ij})$; *K*=number of clusters.

2. From an initial assignment of the items into *K* clusters.

3. Locate the medoid for each cluster.

4a. For *K* medoids clustering:

- For the m^{th} cluster reassign the i_m^{th} to its nearest cluster medoid then the objective function is:

$\mathrm{ESS}_{med} = \sum_{m=1}^{K} \sum_{c(i)=m} dii_m$ is reduced in magnitude, where c(i) is the cluster containing the ith item.

- Repeat step 3 and reassignment step until no further reassignment of items takes place.

4b. For partition around medoids clustering:

- For each cluster, swap the medoid with the non-medoid item gives the largest reduction in ESS_{med}.

Repeat swapping process over all clusters until no reduction in ESS_{med} takes place.

2.5. Data imputation (Kurt, 2012 [17])

Generally the nonhierarchical clustering algorithms are not recommended in the presence of missing values. In the ALL dataset which consists of numerical as well as categorical observations there are missing values. As such it is logical to impute the missing values in this dataset. In the present work Expectation Maximization (EM) algorithm for unrestricted model (Shafer J.L., 1997 [22]; Little and Rubin, 1987 [18]) available in the mix library in R has been used for data imputation in the ALL dataset.

3. Previous studies

Siddheswar and Huri (1999 [23]) overcome the disadvantage of the *K-mean* algorithm which determines the number of clusters, *k*, by developing a simple validity measure based on the intra cluster and inter cluster distance measures that allows the estimation number of clusters automatically minimizing validity measures . Kanungo et al.(2002 [14]) conducted a study to compare the efficiency of *K*-means algorithm with Balanced Iterative Reducing and Clustering using Hierarchies (BIRCH) clustering scheme on real and synthetic data from actual applications in image processing, and found that the efficiency of *K*-means is better than BIRCH algorithm. Chris and He (2004 [5]) conducted a study to test the effect of dimension reduction on *K*-means clustering by using principal component analysis to reduce the data from the original 1000 dimension to 40, 20, 10, 6 and 5 dimensions respectively on 4029 of Gene expression of 96 tissue samples on human Lymphoma. They have applied *K*-means on 10 random samples of each new groups combination [40, 20, 10, 6 and 5 dimensions]. They found that the results systematically and significantly improved. Swami and Jain (2006 [24]) conducted a study to evaluate the accuracy of PAM clustering considering a diabetes dataset containing 786 records with 8 attribute and 2 classes. They implemented the method of PAM on their data and compared the results with other methods of classification which

are popular for classifying diabetics data like C4.5, CBA, CHAR and found that the accuracy of their method is near to the other popular classification methods for diabetics. Boomijia (2008 [2]) conducted a study to compare *K*-means with *K*-Medoids algorithms using experimental runs with hundred random data points and found that *k*-Medoids method is more robust than *K*-means in the presence of noise and outliers and *k*-Medoids algorithm performs effectively for small datasets. Devi et al. (2009 [9]) conducted a study to evaluate *K*-means and Partition Around Medoids algorithm by grid environment using design of experiments and found that *K*-means algorithm overcomes the problem of clustering larger datasets and also it clusters the data faster than Partition Around Medoid. Kumar and Sharma (2009 [16]) implemented the *K-mean*s and PAM algorithm using a sample of Leukemia patients datasets with complexity and a high dimensionality of gene and performed a comparative study of the two algorithms and observed that *K-mean*s algorithm is better than PAM when less number of genes is considered for the study. But as the number of genes is increased the average accuracy of PAM clustering improves over *K-mean* clustering.

4. Results and interpretations

4.1 *K*-means clustering for B-cells and T-cells

Table 1 *K*-means results summary for B-cells and T-cells

	For B-cells		For T-cells	
# clusters	Total within (ESS)	Calinski index	Total within (ESS)	Calinski index
2	1452.746	**47.3798**	416.5442	18.40256
3	1145.591	41.9145	358.3031	12.69729
4	888.5717	44.3072	225.9753	**18.4244**
5	864.1490	34.4187	192.7396	16.78645
Max index		47.3798		18.4244

4.2 Partition Around Medoids (PAM) for B-cells and T-cells

Here PAM has been implemented for $k = 2$ to $k = 5$.

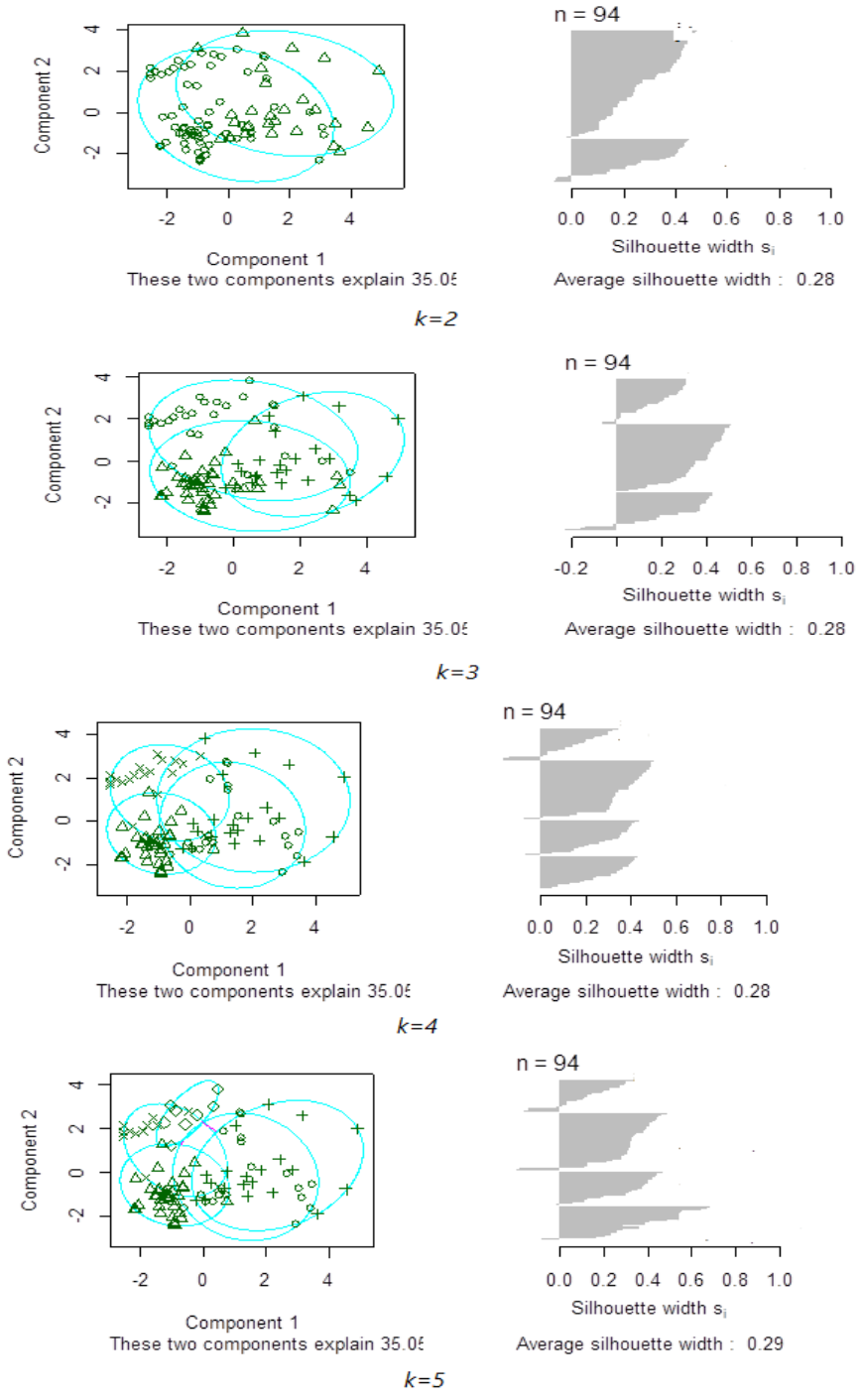

Fig 1 Partition Around Medoids for B-cells. $k = 2$ to 5

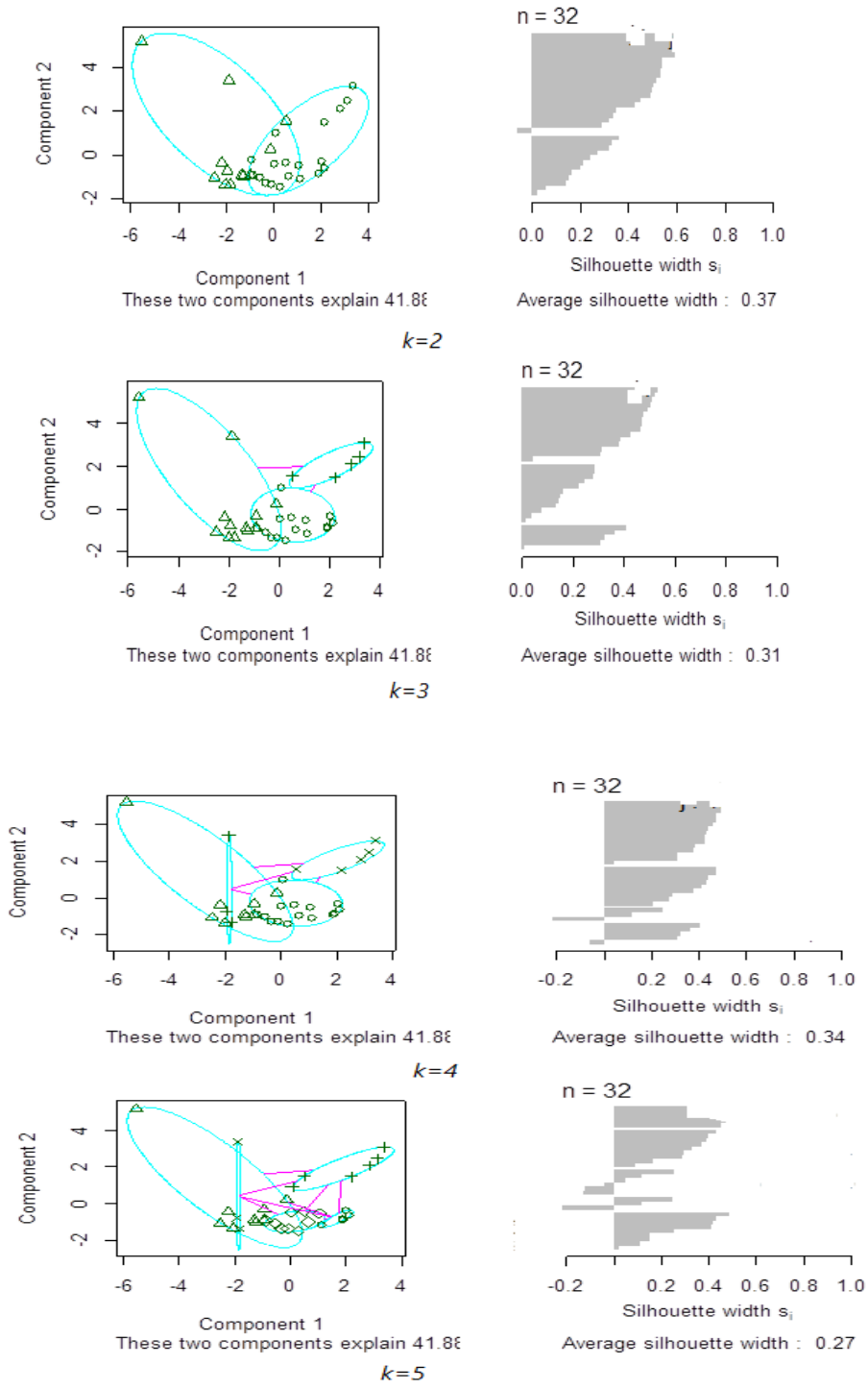

Fig 2 Partition Around Medoids for T-cells. *k* = 2 to 5

Table 2 Average Silhouette width for B-cells and T-cells

	For B-cells				For T-cells			
# clusters	2	3	4	5	2	3	4	5

Average Silhouette width	0.28	0.28	0.28	**0.29**	0.37	0.31	0.34	**0.27**

Table 3 Optimal internal and stability results of *K*-means and PAM clustering for B-cells and T-cells

	B-cells		T-cells	
	Score	Method	Score	Method
Connectivity	22.3202	*K*-means (**2**)	9.7500	PAM (**2**)
Dunn	00.2119	*K*-means (**5**)	0.2967	*K*-means (**5**)
Silhouette	00.3044	*K*-means (**4**)	0.3703	*K*-means (**3**)
APN	00.0372	*K*-means (**4**)	0.0134	*K*-means (**2**)
AD	09.0340	PAM (**5**)	6.8256	*K*-means (**5**)
ADM	00.3642	*K*-means (**2**)	0.2016	*K*-means (**2**)
FOM	00.8503	*K*-means (**5**)	0.7185	*K*-means (**5**)

The bold values between brackets is the optimal number of clusters

Internal validation and stability measures results after reduced dimension

We have also implemented clustering on variables to reduce dimension of the data sets which resulted in the reduction of dimension from 16 variables (BT, sex, remiss, CR, t411, t922, cyton, citog,molb, fusionp, mdr, kinet, ccr, relapse, transp, f.u.) to only 8 variables (BT,remiss, CR,t922, cyton, citog, molb, fusionp).

The results of the internal validation and stability measures of the data after dimension reduction for both data subset are presented in table (4)

Table 4 Optimal internal and stability results of *K*-means and PAM clustering for B-cells and T-cells with dimension reduction

	B-cells		T-cells	
	Score	Method	Score	Method
Connectivity	**4.9619**	*K*-means (**2**)	**7.3702**	*K*-means (**2**)
Dunn	0.1295	*K*-means (**3**)	**0.3343**	*K*-means (**2**)
Silhouette	**0.4115**	*K*-means (**2**)	**0.4570**	*K*-means (**2**)
APN	0.0575	*K*-means (**2**)	0.0467	*K*-means (**2**)
AD	**4.9102**	PAM (**5**)	**3.5676**	PAM (**5**)
ADM	0.4670	*K*-means (**2**)	0.2503	*K*-means (**2**)
FOM	1.0225	PAM (**5**)	0.8802	PAM (**5**)

The bold values between brackets is the optimal number of clusters

5. Concluding Remarks and Recommendations

K-means and Partition Around Medoids (PAM) are both partitioning method and attempt to minimize the distance between objects inside cluster,

The estimated number of clusters using *K*-means with Calinski index (Calinski and Harabasz 1974[4]) for B-cells data has been found to be 2, while for T-cells it is 4 (see table 1).

Using the average Silhouette width for PAM, the estimated number of clusters for B-cells has been found to be 5, while for T-cells it is 2 as in table 2. But here it should be kept in mind that the cluster structures by PAM for B-cells and T-cells are artificial, because average Silhouette width is less than 0.5 (Kaufman and Rousseeuw 2005 [15]).

Internal validation and stability measures have been used to compare between *K*-means and PAM for B-cells and T-cells and the findings have been exhibited in table 3 and it has been found that for B-cells the K-means performs better than PAM by Connectivity, Dunn, Silhouette, APN, ADM, FOM indexes and PAM perform better than K-means by AD index. For T-cells, PAM performs better than K-means by Connectivity index and *K*-means performs better than PAM by Dunn, Silhouette, APN, AD, ADM, FOM indices.

From table (3) and table (4) we can conclude that there is some modifications respected to internal stability (Connectivity, Dunn and Silhouette) for T-cells and (Connectivity and Silhouette) for B-cells also there is some modifications in stability measures respected to AD measures for both B-cells and T-ells.

Therefore the main recommendations from the current investigation are:

I. Internal and stability validation should be used to select the appropriate method for a given data set in clustering analysis.

II. Suitable index should be used for the appropriate method (s)

References

[1] Anja, S., Mia, H. and Rousseeuw, P. (1997) Clustering in an Object-Oriented Environment, *Journal of Statistical Software*, 1 (4), 1-30.

[2] Boomijia, D. (2008) Comparison of Partition Based Clustering Algorithms, *Journal of Computer Applications*, 1 (4), 18-21.

[3] Brito, P., Bertrand, P., Cucumel, G. and Carvalho, F. (2007) *Selected Contributions in Data Analysis and Classification*, Springer, Berlin Hiedelberg, 161-163.

[4] Calinski, T. and Harabasz, J. (1974) A Dendrite Method for Cluster Analysis, *Communications in Statistics*, 3 (1), 1-27.

[5] Chris, D. and Xiaofeng, He. (2004) K-means Clustering via Principal Component Analysis, *21st International Conference on Machine Learning*, Canada, 29-36.

[6] Cluster Analysis: *http: // www.stat.berkeley.edu /~s133/ Cluster2a. html*.

[7] Datta, S. and Datta, S. (2003) Comparisons and Validation of Statistical Clustering Techniques for Microarray Gene Expression Data, *Bioinformatics*, 19 (4), 459-66.

[8] Dean, W. and Richard, J. (2002), *Applied Multivariate Statistical Analysis*, 5th edition, Pearson Education, USA, 694-697.

[9] Devi, B., Ramachandram, S. and Retta, S. (2009-2010) Performance Evaluating of Partition Based Clustering Algorithms in Grid Environment Using Design of Experiments, Computer Science Department, Osmania University, Hyderabad, India, *International Journal of Reviews in Computing* , IJRIC & LLS, 46-53.

[10] Everitt, S. B., Stahl, D., Leese, M. and Landau, S. (2011) *Cluster Analysis*, 5th Ed., Wiley series in Probability and Statistics, UK.

[11] Irizarry, R., Hobbs, B., Collin, F., Beazer-Barclay, D., Antonellis, J. , Scherf, U., Speed, P. (2003) Exploration, Normalization, and Summaries of High Density Oligonucleotide Array Probe Level Data, *Biostatistics*, 4 (2): 64- 249.

[12] Izenman, A. J. (2008) *Modern Multivariate Statistical Techniques, Regression, Classification and Manifold Learning*, Springer Science & Business, Media Philadelphia, PA 19122, USA.

[13] Jonathan, M. G., Daniele, S. and Khhairul, A. R. (2010) Consensus Clustering and Fuzzy Classification for Breast Cancer Prognosis, *24th European 208 Conference on Modeling and Simulation, June 1st - 4th* , Kuala Lumpur, Malaysia, 15-22.

[14] Kanungo, T., Mount, D., Neteanyaho, N., Piatko, C., Silverman, R. and Wu, A. (2002) An Efficient k-means Clustering Algorithm: Analysis and Implementation, *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 24 (7), July, 881-892.

[15] Kaufman, L. and Rousseeuw, P. (2005) *Finding Groups in Data, An Introduction to Cluster Analysis*, John Wiley & Sons Hoboken, New Jersey.

[16] Kumar, H. and Sharma, V. (2011) A Comparative Study of *k*-mean and PAM Algorithms using Leukemia Datasets, *International Symposium on Computing, Communication, and Control, ISCCC , Proc. of CSIT* ,Vol. 1(2011), IACSIT Press, Singapore, 136-140.

[17] Kurt, H. (2012) With Contributions from Walter Boehm *package "clue",* Version 0.3-45 October 16.

[18] Little, R. and Rubin, D. (1987) *Statistical Analysis with Missing Data*, John Wiley & Sons, California, USA.

[19] Qin, H. (1999) A Review of Clustering Algorithms as Applied in IR, *UIUCLIS* (1999) */ 6+IRG*, 1-33.

[20] *R manual documentation* (2012).

[21] Sabina, L., Gentleman, R, Antonella, V., Franco, M., Ritz, J., and Robin, F. (2004) Gene Expression Profile of Adult T-cell Acute Lymphocytic Leukemia Identifies Distinct Subsets of Patients with Different Response to Therapy and Survival, *Blood Journal*, 103 (7), 2771-2778.

[22] Schafer, J. (1997) *Analysis of Incomplete Multivariate Data*, Chapman & Hall / CRC, Florida, USA.

[23] Siddheswar, R. and Rose, H. T. (1999) Determination of Number of Clusters in k-Means Clustering and Application in Colour Image Segmentation, 27-29 December 1999, *Narosa publishing house, ISBN:81-7319-347-9*, 137-143, Calcutta, India.

[24] Swami, D. and Jain, R. (2006) PAMC: Partition Around Medoids for Classification, Information, *Information Technology Journal*, 5(6), 1102-1105.

[25] Yeung, K. Y., Haynor, D. R. and Ruzzo, W. L. (2001) Validating Clustering for Gene Expression Data, *Bioinformatics*, 17(4), 309-18.