Biomedical Statistics and Informatics
2020; 5(1): 20-25
http://www.sciencepublishinggroup.com/j/bsi
doi: 10.11648/j.bsi.20200501.14
ISSN: 2578-871X (Print); ISSN: 2578-8728 (Online)

Review Article

Comparative Study of K-Means, Partitioning Around Medoids, Agglomerative Hierarchical, and DIANA Clustering Algorithms by Using Cancer Datasets

Md. Bipul Hossen[1, *], Md. Rabiul Auwul[2]

[1]Department of Statistics, Begum Rokeya University, Rangpur, Bangladesh

[2]Department of Statistics, Guangzhou University, Guangzhou, China

Email address:

mbipu@brur.ac.bd (Md. B. Hossen), rabiulauwul@gmail.com (Md. R. Auwul)

[*]Corresponding author

To cite this article:

Md. Bipul Hossen, Md. Rabiul Auwul. Comparative Study of K-Means, Partitioning Around Medoids, Agglomerative Hierarchical, and DIANA Clustering Algorithms by Using Cancer Datasets. *Biomedical Statistics and Informatics.* Vol. 5, No. 1, 2020, pp. 20-25.
doi: 10.11648/j.bsi.20200501.14

Received: December 29, 2019; **Accepted**: January 10, 2020; **Published**: March 2, 2020

Abstract: Clustering plays a particularly fundamental role in exploring data, creating predictions and to overcome the anomalies in the data. Clusters that contain parallel, identical characteristics in a dataset are grouped using reiterative algorithms. As the data in real world is rising day by day so the challenges of perceiving and interpreting the consequential mass of data, which often consists of millions of measurements are increased by the intricacy of a huge number of genes of biological networks. To addressing this challenge, we use clustering algorithms. In this study, we provided a comparative study of the four most popular clustering algorithms: K-Means, PAM, Agglomerative Hierarchical and DIANA and these are evaluated on eight real cancer (four Affymetrix and four cDNA) gene data and simulated data set. The comparative results based upon seven popular cluster validity indices: Average Silhouette Index, Corrected rand Index, Variation of Information, Dunn Index, Calinski-Harabasz Index, Separation Index, and Pearson Gamma. We determine that PAM is best for Affymetrix data set and DIANA is best for cDNA dataset among these four clustering algorithms. This study provides practical evaluation frameworks for accessing clustering results on gene expression cancer datasets.

Keywords: Microarray, Clustering Algorithm, Gap Statistic, Validity Indices

1. Introduction

Microarrays technology can concurrently measures the thousands of genes expression level within a particular mRNA biological sample and across collections of all related samples [1]. Such technology can be used to compare the level of gene expression in order to identify diagnostic or prognostic genes, classify genes, and monitor the response to therapy. For these reasons, microarrays technology are considered important tools for discovery in the medical community. A large number of genes and the complexity of biological networks greatly increase the challenges of comprehending and interpreting the resulting mass of data, which often consists of millions of measurements [2]. A first step toward addressing this challenge is the use of clustering techniques, which is essential in the data mining process to reveal natural structures and identify interesting patterns in the underlying data.

In data mining, there are two learning approaches- Supervised and Unsupervised learning. Clustering is unsupervised learning and defined as it is the task of grouping a set of observations into subsets (called clusters) so that observations in the same cluster are similar in some sense. Clustering techniques have extensively contribute in the various fields including, artificial intelligence, pattern recognition, bioinformatics, segmentation and machine learning. An appropriate Clustering algorithm is highly demanded to extract hidden information from co-expression analysis of enormous genome data [3]. In that case, a

common task is to compare the Clustering algorithms for gene expression datasets.

Generally single channel microarrays (Affymetrix) and double channel microarrays (cDNA) are two types of platforms where the gene expression microarray technology is existing and these datasets are significant to cluster both genes and samples [4, 5]. The above types of datasets are usually used for gene based clustering and sample based clustering. The sample based clustering only conducted in this study. And in sample based clustering genes are treated as features while samples are treated as objects and samples are partitioned into homogeneous groups.

There are numerous broadly used Clustering algorithms are already developed to capture the overall feature of high dimensional variable datasets. K-Means [6], Partitioning Around Medoids (PAM) [7], Agglomerative Hierarchical methods [8] and Divisive Analysis Methods (DIANA) [9] are more popular between them. Therefore this paper performs a comparative analysis of above four clustering algorithms. The performance of theses clustering algorithms is compared in terms of accuracy and efficiency through seven validity indices [10] (Average Silhouette Width, Corrected Rand Index, Variation of Information, Dunn Index, Calinski-Harabasz Index, Separation Index and Pearson Gamma). Since in this study, some packages (cluster, clusterCrit, clusterSim, limma, fpc and ggplots2, Clus_Stat etc.) were used with R 3.2.5 version.

2. Materials and Methods

2.1. The Gap Statistic

The gap statistic [11] is used for finding an optimal number of clusters (K) in a dataset and also gives the idea behind their approach was to find a way to standardize the comparison of $Log\ W_k$ with a null reference distribution of the data, i.e. a distribution with no obvious clustering. Their estimate for the optimal number of clusters k is the value for which $Log\ W_k$ falls the farthest below this reference curve. This formula for calculating the gap statistic is:

$$Gap_n(k) = E_n \times (Log\ W_k) - Log\ W_k$$

Where E_n denotes the expectation under n sample size from the reference distribution. The estimated will be the value maximizing $Gap_n(k)$ after taking the sampling distribution into account.

2.2. Algorithms

2.2.1. The K-Means Algorithm (KM)

The k-means algorithm [6] is one of the simplest unsupervised learning algorithms to classify a given data set through a certain number of clusters (assume k clusters) static a priori. To decrease the complexity of grouping data it can be run multiple times. How this algorithm works that are explained in Figure 1.

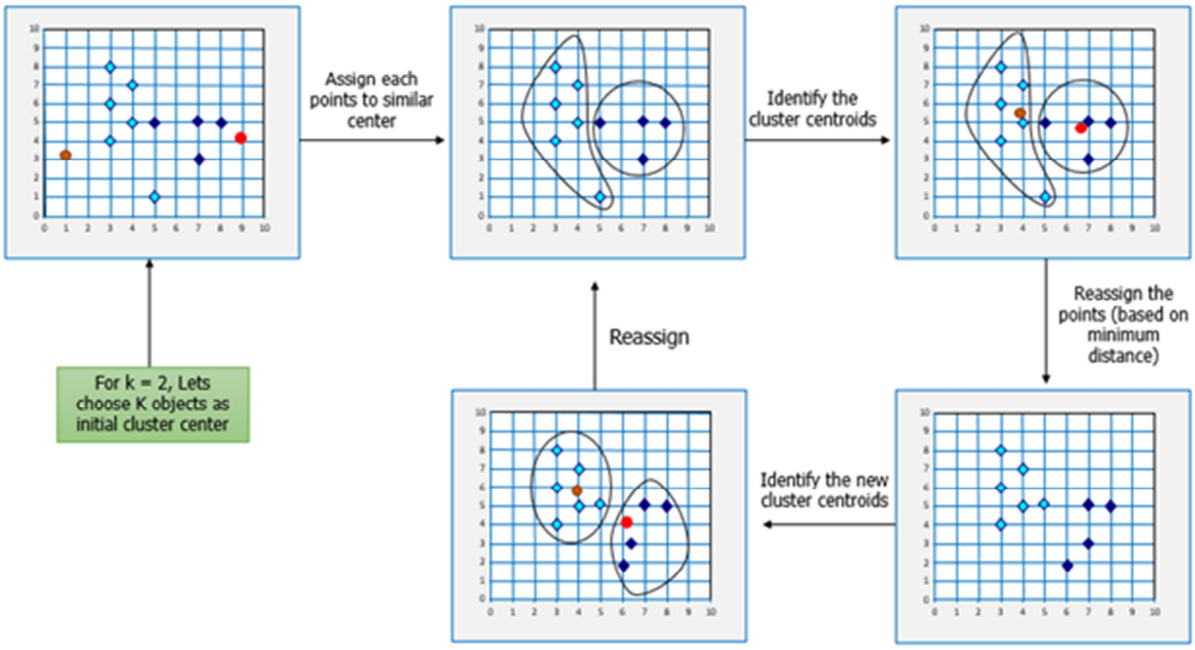

Figure 1. *Principle flow of k-means algorithm.*

2.2.2. The Partitioning Around Medoids (PAM) Algorithm

The k-means algorithm is considerate to outliers because an object with exceptionally large value may substantially change the distribution of data. In this algorithm, a medoid can be used instead of the mean value of compelling the objects in a cluster which is the most centrally located object in a cluster. Based on the standard of reducing the sum of the differences between each object and its consistent reference point can still be performed as the partitioning method and this forms can perform on the basis of k-Medoids and it is called. Partitioning Around Medoids [7]. The basic strategy of PAM clustering algorithms is to find k clusters in n objects by first randomly judgment a representative object (the

medoids) for each cluster are showed in Figure 2.

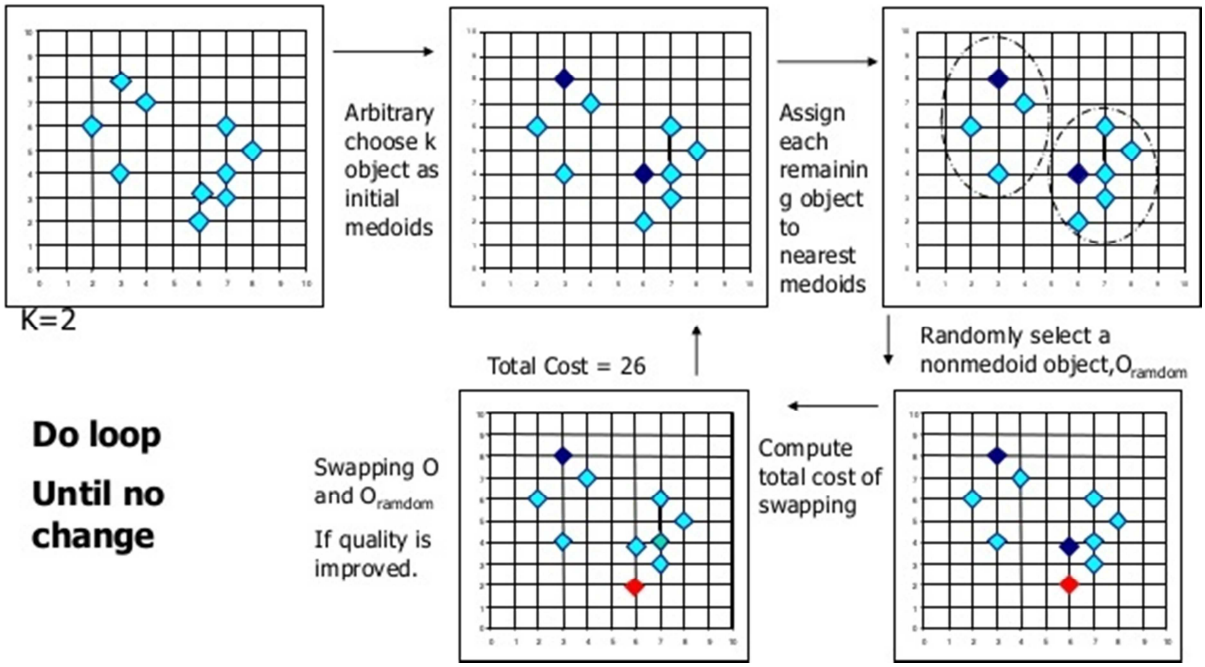

Figure 2. *Principle flow of PAM Algorithm.*

2.2.3. Agglomerative Hierarchical Clustering Algorithm (AHC)

Agglomerative hierarchical clustering or bottom-up clustering method start with each object presenting a cluster, and then the methods gradually merge theses clusters into large ones [8, 12]. These algorithms start with each object presenting a cluster, and then the methods gradually merge theses clusters into large ones. For each of the successive iteration it agglomerates (merges) the closest pair of clusters by satisfying some similarity criteria, until all of the data is in one cluster that are clarified in Figure 3.

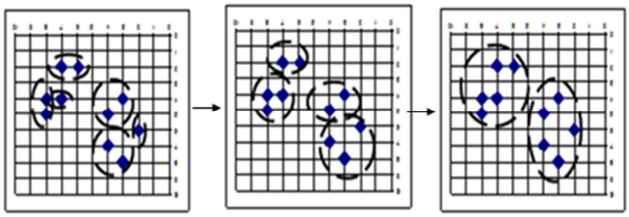

Figure 3. *Principle flow of AHC Algorithm.*

2.2.4. Divisive Analysis Clustering (DIANA)

Divisive Analysis Clustering [9] is a hierarchical clustering technique which constructs the hierarchy in the inverse order and this approaches is the reversal algorithm of Agglomerative Hierarchical Clustering. One larger cluster consisting of all n objects split into two clusters until finally all clusters, comprise of single objects which is illustrated in Figure 4.

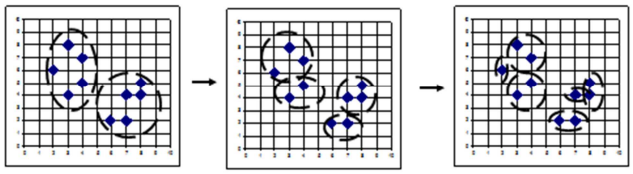

Figure 4. *Principle flow of DIANA Algorithm.*

2.3. Clustering Validity Indices

Cluster validity indices [10] are functions that help a user answer the question of whether a particular clustering of the data is better than an alternative clustering. For unsupervised clustering, where partitions are made without reference to external classes, these cluster validity metrics must rely only on internal measures of the data. Several such validity metrics exist, such as within-cluster distances (should be low) and between-cluster distances (should be high). Several cluster validity indices are briefly discussed in Table 1.

Table 1. *Short Description of Validity Indices.*

Validity Indices	Functions	Descriptions
Average Silhouette Width (ASW)	$\frac{1}{m}\sum_{i=1}^{m}\frac{v(i)-\mu(i)}{max\{v(i),\mu(i)\}}$	Its values within [−1, 1]. The optimal value is the highest.
Corrected Rand Index (CRI)	$\frac{\sum_{i,j}\binom{n_{ij}}{2}-\left[\sum_{i}\binom{a_i}{2}\sum_{j}\binom{b_j}{2}\right]/\binom{n}{2}}{\frac{1}{2}\left[\sum_{i}\binom{a_i}{2}\sum_{j}\binom{b_j}{2}\right]-\left[\sum_{i}\binom{a_i}{2}\sum_{j}\binom{b_j}{2}\right]/\binom{n}{2}}$	Its values within [−1, 1]. The optimal value is the highest.
Variation of Information (VI)	$[H(C)-I(C,C/)]+[H(C/)-I(C,C/)]$	Its values within [0, 1]. The optimal value is the lowest.

Validity Indices	Functions	Descriptions
Dunn Index (DI)	$\min\{\min\{\frac{d(i,j)}{\max d'(k)}\}\};1 \le i, j \le n;$	Its values within [0, 1]. The optimal value is the highest.
Calinski-Harabasz Index (CH)	$[trace(SB)/trace(Sw)].[(np-1)/(np-k)]$	Its values within [0, ∞]. The optimal value is the highest.
Separation Index (SI)	$\frac{\sum_{i=1}^{m}\sum_{j=1}^{n} u_{ij}^{2} d(Xj - Vi)^2}{N * (d\ min)^2}$	Its values within [0, ∞]. The optimal value is the lowest.
Pearson Gamma (PG)	p(d; m), d *is the* divergence vector, m is binary vector	Its values within [0, 1]. The optimal value is the highest.

2.4. Data Sets

The datasets present different values for features such as type of microarray chip (second column), tissue type (third column), number of samples (fourth column), number of classes (fifth column), number of samples within the classes (sixth column), dimensionality (seventh column) and (last column) shows the dimensionality after feature selection. Short description of these datasets in are presented in Table 2.

Table 2. *Short description of Cancer Data Sets.*

Name of Datasets	Chip	Tissue	N	C	Dist. Classes	M	d
Chowdary [13]	Affy	Breast	104	2	62, 42	22283	182
Pomeroy-V1 [14]	Affy	Brain	34	2	25, 9	7129	857
Golub-V2 [15]	Affy	Bone marrow	72	3	38, 9, 25	7129	1877
Nutt-V1 [16]	Affy	Brain	50	4	14, 7, 14, 15	12625	1377
Bittner [17]	cDNA	Skin	38	2	19, 19	8067	2201
Risinger [18]	cDNA	Endometrium	42	4	13, 3, 19, 7	8872	1771
Tomlins-V2 [19]	cDNA	Prostate	92	4	27, 20, 32, 13	20000	1288

3. Results and Discussions

3.1. Simulated Data Analysis

To check the performance of clustering method it introduced a simulated data set that has 150 rows as genes and 8 columns as sample. First 1-50 gene are high1y expressed, 51-100 gene are medium expressed and last 101-150 gene present low expressed in terms of intensity level. The simulated data are generated from normal distribution N (5, 12). Therefore we introduce three cluster as three main effect.

Figure 5 represents gap statistic and observed that when the number of cluster is 3 than the Gap statistic gives the optimal value. Therefore we may conclude that three clusters are presented in the simulation data.

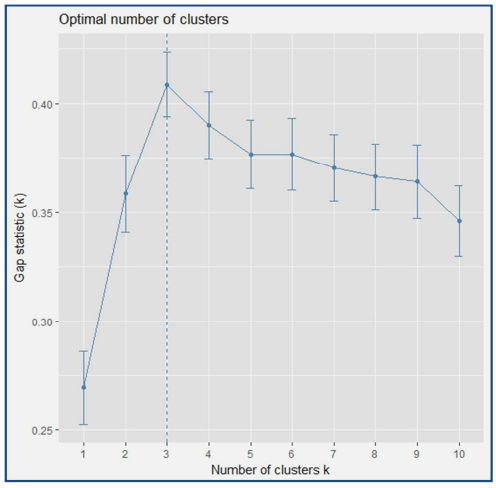

Figure 5. *Gap Statistic of Simulated data set.*

Table 3. *Average validation score of simulation data for different clustering methods.*

Alg.	ASW	CRI	VI	DI	CH	PG	SI	N*
KM	0.346469	0.960277	0.112456	98.62022	0.254336	1.58735	0.625650	4
PAM	0.337935	0.920752	0.243079	95.88044	0.211840	1.41413	0.692231	2
AHC	0.346063	0.941045	0.152512	98.42486	0.254336	1.55375	0.702224	2
DIANA	0.346469	0.960263	0.112456	98.62012	0.254336	1.58735	0.625550	3

[N*=Total number of Optimal Indices].

The analysis of the simulation data result presented in Table 3 and we see that there are maximum numbers of validity indices satisfied by K-Means followed by DIANA clustering algorithms. So we can say that K-means and DIANA are the best clustering methods than PAM and Agglomerative Hierarchical algorithm for simulated data.

3.2. Comparative Results of K-means, PAM, AHC and DIANA for Affymetrix Datasets

We applied the all clustering methods to the 4 set of affymetrix real datasets and also check their accuracy through several indices are given in Table 4 along with the

graphical technique as in Figure 6.

Table 4. *Average validation score of Affymetrix datasets for different clustering methods.*

Alg.	ASW	CRI	VI	DI	CH	PG	SI	N*
KM	0.305645	0.12716	1.04765	0.46476	19.3129	0.64567	35.4125	1
PAM	0.169554	0.22672	1.04201	0.44581	6.33014	0.34439	31.9314	3
AHC	0.304799	0.12976	1.04217	0.45420	19.2508	0.64596	35.2290	1
DIANA	0.406875	0.03670	0.96125	0.76425	9.84709	0.60761	55.1780	2

[N*=Total number of Optimal Indices].

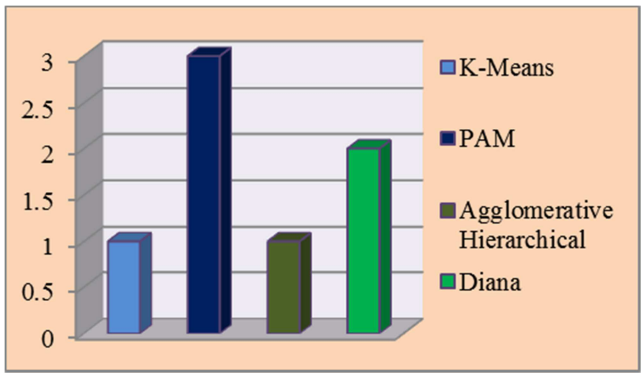

Figure 6. *Bar plot of the total number of optimal indices satisfied by different clustering methods for Affymetrix data.*

Table 4 demonstrate that the comparative analysis of four clustering algorithm and this analysis will evaluate the several measurements of indices. For K-Means and Ag. Hierarchical clustering we see only one optimal index were performed better. In DIANA clustering algorithm there are two indices performed better but in PAM clustering algorithm we see there are three indices performed better. Maximum numbers of validity indices satisfied by PAM clustering algorithm. Figure 6 also represents the comparative analysis and it shows that the maximum number of optimal indices happened in PAM clustering algorithm among others. Therefore we may conclude that PAM clustering algorithm is the best followed by DIANA, K-Means, and Agglomerative Hierarchical methods for Affymetrix datasets.

3.3. Comparative Results of K-means, PAM, AHC and DIANA for cDNA Datasets

We applied all clustering methods to the 4 set of cDNA real datasets and check their accuracy through several indices are given in Table 5 along with the graphical technique as in Figure 7.

Table 5. *Average validation score of cDNA datasets for different clustering methods.*

Alg.	ASW	CRI	VI	DI	CH	PG	SI	N*
KM	0.101433	0.092254	1.835807	0.441071	**7.722119**	0.419621	35.01233	1
PAM	0.082616	0.105567	1.660156	0.393937	6.480427	0.362082	33.00171	1
AHC	0.086476	0.118176	1.687551	0.418354	7.231679	0.344757	35.636	1
DIANA	0.139261	0.027782	1.613097	0.509219	6.275163	0.554526	40.2396	4

[N*=Total number of Optimal Indices].

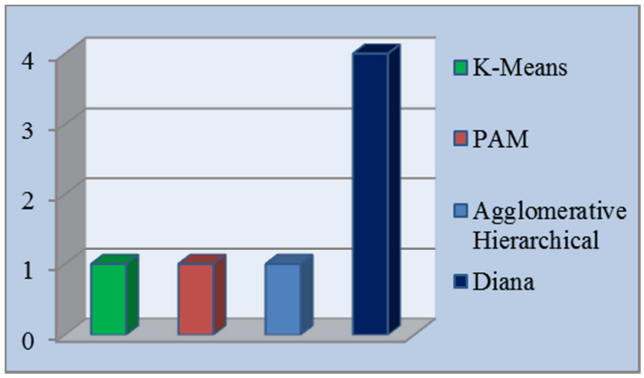

Figure 7. *Bar plot of the total number of optimal indices satisfied by different clustering methods for cDNA data.*

Table 5 shows that the comparative analysis of four clustering algorithm according to the several measurements of indices. We observed that for DIANA clustering algorithm maximum number of validity indices performed better but in others clustering algorithm only one indices performed better. Figure 7 also represents the comparative analysis and it shows that the maximum number of optimal indices happened in DIANA clustering algorithm among others. Therefore we may conclude that DIANA clustering algorithm is the best algorithm among the others for cDNA datasets.

4. Conclusions

Cluster analysis problem has always interested scientists as it deals with the grouping of objects having common properties and it run as a first step of data summary and grouping genes in a microarray gene expression data analysis. As we show here a comparative study of four clustering algorithms applied on the simulated data and eight clinical cancer gene expression datasets. Our results reveal that, K-means and DIANA clustering methods perform well for simulated data. The PAM gives the best performance for Affymetrix datasets. For cDNA datasets, the DIANA clustering exhibited the best performance in terms of recovering the true structure of the datasets. To the best of our knowledge, the comparative study of K-means, PAM,

Agglomerative Hierarchical clustering and DIANA with several validity indices as Average Silhouette Width, Corrected rand Index, Variation of Information, Dunn Index, Calinski-Harabasz Index, Separation Index, and Pearson Gamma are poorly documented in the literature.

Acknowledgements

Declared none.

References

[1] Schena M., Shalon D., Davis R. W., Brown P. O. (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray, Science, 270, 467–470.

[2] Hossen M. B. Siraj-Ud-Doulah M. Hoque M. A. (2015) Methods for Evaluating Agglomerative Hierarchical Clustering for Gene Expression Data: A Comparative Study, Computational Biology and Bioinformatics, 3 (6), 88-94.

[3] Hossen M. B., Mowla A., Rashid or H., Binyamin M. (2017) On the Selection of Appropriate Proximity Measurement for Gene Expression Data, International Journal of Biomedical Materials Research, 5 (5), 59-63.

[4] Daxin J., Chun T., Aidong Z. (2004) Cluster Analysis for Gene Expression Data: A Survey, IEEE Transactions on Knowledge and Data Engineering, 16 (11), 1370-1386.

[5] Costa I. G., Carvalho F. A. D., Souto M. C. P. D. (2004) Comparative Analysis of Clustering Methods for Gene Expression Time Course Data, Genetics and Molecular Biology, 27 (4), 623-631.

[6] MacQueen J. B. Some methods for classification and analysis of multivariate observations. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, University of California Press 1967, 1, 281-297.

[7] Kaufman L., Rousseeuw P. J. Finding Groups in Data: An Introduction to Cluster Analysis. Wiley, New York, 1990.

[8] Hossen M. B., Siraj-Ud-Doulah M. (2017) Identification of robust clustering methods in gene expression data analysis, Current Bioinformatics, 12 (6), 558-562.

[9] Patnaik A. K., Bhuyan P. K., Krishna R. K. V. (2016) Divisive Analysis (DIANA) of hierarchical clustering and GPS data for level of service criteria of urban streets, Alexandria Engineering Journal, 55 (1), 407-418.

[10] Arbelaitz O., Gurrutxaga I., Muguerza J., Pérez J. M., Perona I. (2013) An extensive comparative study of cluster validity indices, Pattern Recognit, 46, 243–256.

[11] Tibshirani R., Walther G., Hastie R. (2001) Estimation the number of cluster in a data via gap statistic, J. R. Statist. Soc. B, 63 (2), 411-423.

[12] Jain A. K., Dubes R. C. Algorithms for clustering data, Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1988.

[13] Chowdary D., Lathrop J., Skelton J., Curtin K., Briggs T., Zhang Y., Yu J., Wang Y., Mazumder A. (2006) Prognostic gene expression signatures can be measured in tissues collected in RNA later preservative, J Mol Diagn, 8, 31–39.

[14] Pomeroy S. L., Tamayo P., Gaasenbeek M., Sturla L. M., Angelo M., McLaughlin M. E., Kim J. Y., Goumnerova L. C., Black P. M., Lau C., Allen J. C., Zagzag D., Olson J. M., Curran T., Wetmore C., Biegel J. A., Poggio T., Mukherjee S., Rifkin R., Califano A., Stolovitzky G., Louis D. N., Mesirov J. P., Lander E. S., Golub T. R. (2002) Prediction of central nervous system embryonal tumour outcome based on gene expression, Nature, 415, 436-42.

[15] Golub T. R., Slonim D. K., Tamayo P., Huard C., Gaasenbeek M., Esirov J. P., Coller H., Loh M. L., Downing J. R., Caligiuri M. A., Bloomfield C. D., Ander E. S. (1999) Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring, Science, 286 (5439), 531-537.

[16] Nutt C. L., Mani D. R., Betensky R. A., Tamayo P., Cairncross J. G., Ladd C., Pohl U., Hartmann C., McLaughlin M. E., Batchelor T. T., Black P. M., von Deimling A., Pomeroy S. L., Golub T. R., Louis D. N. (2013) Gene expression-based classification of malignant gliomas correlates better with survival than histological classification, Cancer Research, 63 (7), 1602-1607.

[17] Bittner M., Meltzer P., Chen Y., Jiang Y., Seftor E., Hendrix M., Radmacher M., Simon R., Yakhini Z., Ben-Dor A., Sampas N., Dougherty E., Wang E., Marincola F., Gooden C., Lueders J., Glatfelter A., Pollock P., Carpten J., Gillanders E., Leja D., Dietrich K., Beaudry C., Berens M., Alberts D., Sondak V., Hayward N., Trent J. (2000) Molecular classification of cutaneous malignant melanoma by gene expression profiling, Nature, 406, 536-540.

[18] Risinger J. I., Maxwell G. L., Chandramouli G. V. R., Jazaeri A., Aprelikova O., Patterson T., Berchuck A., Barrett J. C. (2013) Microarray analysis reveals distinct gene expression profiles among different histologic types of endometrial cancer, Cancer Research, 63, 6–11.

[19] Tomlins S. A., Mehra R., Rhodes D. R., Cao X., Wang L., Dhanasekaran S. M., Kalyana-Sundaram S., Wei J. T., Rubin M. A., Pienta K. J., Shah R. B., Chinnaiyan AM. (2007) Integrative molecular concept modeling of prostate cancer progression, Nature Genetics, 39, 41-51.

[20] Khan J., Wei J. S., Ringner M., Saal L. H., Ladanyi M., Westermann F., Berthold F., Schwab M., Antonescu C. R., Peterson C., Meltzer P. S. (2001) Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks, Nat Med, 7 (6), 673–679.