The Use of Inductive Methods to Identify Subtypes of Glioblastomas in Gene Clustering

Iryna Lurie [1[0000-0001-8915-728X]], Volodymyr Lytvynenko[1[0000-0002-1536-5542]], Serge Olszewski[2[0000-0003-4499-8485]], Mariia Voronenko[1[0000-0002-5392-5125]], Alexander Kornelyuk[3[0000-0003-0146-2832]], Ulzhalgas Zhunissova[4[0000-0001-5255-9314]] and Oleg Boskin [1[0000-0001-7391-0986]]

[1]Kherson National Technical University, Kherson, Ukraine,
[2] Taras Shevchenko National University of Kyiv, Kyiv, Ukraine,
[3]Institute of Molecular Biology and Genetics, Kyiv, Ukraine,
[4]Astana Medical University, Astana, Kazakhstan
lurieira@gmail.com, immun56@gmail.com,
olszewski.serge@gmail.com,mary_voronenko@i.ua,
kornelyuk@imbg.org.ua, Ulzhalgaszhunisova@gmail.com,
aandre.lenoge@gmail.com

Abstract. The article presents an inductive clustering model of RNA-seq data for solving the problem of identifying glioblastomas subtypes by inductive methods based on k- and c-means algorithms. Comparative studies between inductive and classical iterative clustering algorithms are carried out using the criteria for evaluating clustering and data visualization. The basic principles of creating an inductive model of objective clustering are formed, the ways and prospects of the possible implementation of the model are shown, the advantages of the objective clustering model in comparison with traditional methods of data clustering are determined.

Keywords: Inductive Modeling, Multiform Glioblastoma, Clustering of Biologist Objects, the Method of Group Accounting of Arguments, K-Means Algorithm, External Balance Criterion.

1 Introduction

Glioblastoma multiforme (Glioblastoma multiforme, GBM) is one of the most common and most aggressive types of brain cancer [1] and the leading cause of death in adult brain tumors. Glioblastoma accounts for 52% of all brain tumors. According to the classification of central nervous system tumors by the World Health Organization, the standard term for this brain tumor is “glioblastoma,” and it has two forms: giant cell glioblastoma and gliosarcoma. Glioblastomas are also important brain tumors. It has a very poor prognosis, despite the existence of many therapeutic methods, including surgical resection of a larger tumor volume, followed by concomitant or subsequent chemoradiotherapy. Despite advances in the genomics and classification of

glioma subtypes [2-4], glioblastoma has a worse prognosis than any other cancer of the central nervous system, with an average lifespan of 14 months.
While genomic data continues to grow rapidly, clinical use and treatment transfer are lagging behind. Big data currently stored on the "The Cancer Genome Atlas (TCGA)" network provides a window for creating new clinical hypotheses [5].

One of the main topics is how genomics can be used to obtain clinically relevant information to improve therapy for patients. Two techniques are currently relevant for the formation of an array of gene expression: DNA microarray sequencing method [6] and RNA molecules [7].

The use of the RNA-sequencing method (RNA-seq method) allows you to get the number of studied genes for the studied samples directly. For this reason, this method is more accurate than the DNA microarray method. The number of genes determines the level of activity of this gene or its expression. At the next stage, the problem of identifying the boundary value arises, which allows us to divide the genes into low-expressive and highly expressive. Data needs to be normalized. This involves converting the count values to the same suitable range.

Numerous studies have shown that RNA sequencing technology is more efficient than DNA microarray technology in terms of the quality of the data obtained [8]. Identifying cancer subtypes is an important component of a personalized medicine system. Identifying cancer subtypes is critical when choosing the right treatment for patients, as different subtypes of cancer can respond well to different treatments. Currently, a greater number of computational methods have been developed to detect subtypes of cancer. However, existing methods rarely use information from gene regulation networks to facilitate the identification of subtypes [9]. One of the computational methods that allows us to solve this problem is clustering. Clustering methods are divided into hierarchical and iterative [10].

Hierarchical algorithms are associated with the construction of dendrograms. In agglomerative algorithms, before the start of clustering, all objects are considered separate clusters, which are combined during the algorithm.

However, the hierarchical cluster analysis procedure is good for a small number of objects and is not suitable for large data due to the complexity of the agglomerative algorithm and too large dendrograms. In iterative algorithms, the data is immediately divided into several clusters, the number of which is estimated based on the conditions. Further, the elements are moved between clusters so that a certain criterion is optimized, for example, variability within the clusters is minimized [11].

However, iterative clustering algorithms, in particular, k-means, have several disadvantages:

• It is not guaranteed to achieve the global minimum of the total quadratic deviation, but only one of the local minimums.

• The result depends on the choice of the initial centers of the clusters; their optimal choice is unknown.

• The number of clusters must be known in advance.

High subjectivity is one of the key shortcomings of existing iterative algorithms. Increasing the objectivity of clustering is possible through the use of inductive methods for modeling complex systems based on the inductive method of data processing

[12], in which data processing is carried out by two equal power subsets, and the final decision on the nature of the separation of objects into clusters is made on the basis of integrated use external relevance criteria and internal criteria for assessing the quality of clustering. Thus, the development of models and methods for clustering objects based on inductive modeling methods to solve the problem of identifying cancer subtypes is an urgent task.

2 Problem Statement

A block diagram of the identification of experimental data obtained by RNA sequencing with glioblast is presented in Figure 1.

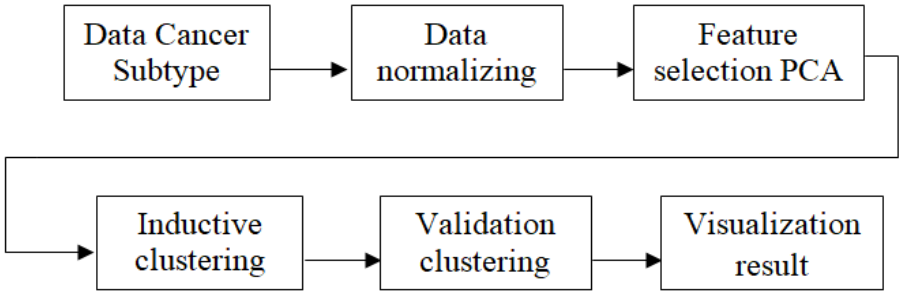

Fig. 1. Procedure for the identification of experimental data obtained by RNA sequencing with glioblastoma.

The paper implements the principles of inductive modeling within the framework of inductive clustering, which suggest the following steps [13]:

- normalization of signs of the studied objects, i.e. their reduction to the same range with one median of the attributes of objects;
- dividing the original data set into two equal in power subsets;
- determination of an external criterion or group of relevance criteria for choosing the optimal clustering for two subsets of the same power;
- selection or development of a basic clustering algorithm used as a component of an inductive model of objective clustering of objects.

Solving the problem of identifying a cancer subtype consists of three main steps: data pre-processing and selection of characteristics, cancer subtype identification methods, verification of results and visualization.

In most cases, genomic data sets are multidimensional and contain noise and missing values. Reducing the size of an element is necessary to remove unnecessary elements and reduce interference. In this paper, we implement the Principal Component Analysis (PCA).

In this paper, four clustering methods are used:

a) classical k-means clustering methods and its fuzzy version of c-means;
b) inductive k-means clustering methods and its fuzzy version of c-means.

For all four approaches, the application of all the indicated clustering procedures both on the initial data and on the data matrix after the Feature Selection procedure using the PCA. Evaluation of the results using Index Dunn, Index Calinski-Harabasz, Entropy, and graphical visualization using Silhouette

The aim of the work is to develop inductive models of object clustering of subtypes by multiglioblastomas based on k- and c-srenich algorithms and to assess the quality of the solution of the results obtained.

3 Review of the Literature

The basic concepts for creating an inductive method for clustering objects are described in [12]. Further development of this theory is reflected in [14]. The concept of objective cluster analysis is presented in the following sections and was further developed in [15]. The authors determine the basic principles of creating an objective cluster inductive model, show the ways and prospects of its implementation, determine the advantages of a cluster inductive model in comparison with traditional methods of data clustering.

Theoretical developments on the implementation of billisterization methods for systems of inductive modeling of complex processes are presented in [16]. In the work [13] authors presented an inductive model of object clustering of objects based on k-means clustering. An algorithm is proposed and practically implemented for dividing the source data into two equal-sized subsets. The paper presents studies on the assessment of the stability of the model to the noise component using the "Seeds" data. However, it should be noted that, despite the successful results achieved in this area, an objective cluster model based on the analysis of cluster systems does not currently have practical implementation for solving problems in bioinformatics.

4 Materials and Methods

4.1 Data

A glioblastoma (GBM) gene expression dataset downloaded from TCGA. This is a small dataset with 1500 genes and 100 cancer samples extracted from gene expression data for examples [5].

4.2 Normalization

Data normalization was carried out according to the characteristics in accordance with the formula:

$$x'_{ij} = \frac{x_{ij} - med_j}{\max(|x_{ij} - med_j|)} \quad (1)$$

where x_{ij} is the value of the attribute i in column j, x'_{ij} is the normalized value of this attribute, med_j is the median of column j. The choice of this normalization method was determined by the fact that as a result, the set of data attributes in all columns had

the same median with a maximum range of variation of attributes from -1 to 1, while the data volume for each column falling into the interquartile distance (50%) is the largest compared to other normalization methods.

4.3 Splitting Into Equidistant Sets

The algorithm for dividing the original set of objects Ω into 2 equally powerful disjoint subsets Ω^A and Ω^B consists of the following steps [16]:

1. calculation of $n \cdot (n-1)2$pairwise distances between objects in the original data sample;
2. selection of a pair of objects $X_s, X_p$the distance between which is minimal:

$$d(X_s, X_p) = \min_{i,j} d(X_i, X_j)$$

3. distribution of the object X_s into a subset Ω^A, and the object X_p into a subset Ω^B; repeating steps 2-3 for the remaining objects. If the number of objects is odd, the last object is distributed into both subsets.

4.4 Inductive *k*-means Algorithm

The k-means algorithm is one of the machine learning algorithms that solves the clustering problem. This algorithm is a non-hierarchical, iterative clustering method; it has gained great popularity due to its simplicity, visualization of implementation, and rather high quality of work. It was invented in the 1950s by the mathematician Hugo Steinhaus [17] and almost simultaneously by Stuart Lloyd [18]. Particularly popular after the publication of the work of McQueen [19] in 1967.

The algorithm is a version of the EM algorithm, which is also used to separate a mixture of Gaussians. The main idea of the k-means algorithm is that the data is randomly divided into clusters, after which the center of mass for each cluster obtained in the previous step is iteratively recalculated, then the vectors are divided into clusters again according to which of the new centers is closer in selected metric.

The purpose of the algorithm is to divide n observations into k clusters so that each observation belongs to exactly one cluster located at the smallest distance from the observation.

Step 1. Start

Step 2. Formation of the initial set Ω of studied objects. Presentation of the data in the form of a matrix $\Omega = \{x_{ij}\}; i = \overline{1,m}$, where n is the number of rows or the number of objects under investigation, m is the number of columns or the number of features characterizing the objects.

Step 3. Data preprocessing - data normalization:

• median normalization (Feature Median) is obtained by calculating the median of all data attributes:

$$z_{ij} = (x_{ij} - med_j) / mad_j$$

where $x_{ij}\left(z_{ij}\right)$ is the i-th observation in the j-th variable (the i-th normalized observation in the j-th variable), $med_j = \underset{i}{med}\left(x_{ij}\right)$ is the median for the j-th variable, $mad_j = \underset{i}{mad}\left(x_{ij}\right)$ is the mean absolute deviation for the j-th variable.

• normalization using a standardized score (z-score) is a measure of the relative spread of the observed or measured value, which shows how many standard deviations is its spread of the relative average value. This is a dimensionless statistic used to compare values of different dimensions or a measurement scale.

$$z_{ij} = \frac{x_{ij} - \overline{X}}{S_{x_{ij}}}$$

where $\overline{X}$ is the average value, $S_{x_{ij}}$ is the standard deviation of the i-th observation in the j-th variable. The best normalization method depends on the data that will be normalized. Typically, the Z-score is very common to normalize the data [20].

Step 4. Dividing Ω into two equally powerful subsets in accordance with the above algorithm. The resulting subsets Ω^A and Ω^B can be formally represented as follows:

$$\Omega^A = \left\{x_{ij}^A\right\};\ \Omega^B = \left\{x_{ij}^B\right\};$$

$$i = \overline{1, n_A} = n_B;\ n_A + n_B = n;\ j = \overline{1, m}$$

Step 5. Choosing the initial number of clusters $k = k_{\min}$.

Step 6. Configuring the k-means clustering algorithm.

For each equidistant subset:

Step 7. Sequential clustering and cluster fixing.

Step 8. Calculation of the internal criteria for the quality of clustering.

Silhouette: $$SWC = \frac{1}{K}\sum_{j=1}^{K} S_{x_j}$$

Dunn Index: $$DI\left(k\right) = \min_{i \in k}$$

Calinski – Harabasz Index: $$QC_{CH} = \frac{QCB \cdot \left(N - K\right)}{QCW \cdot \left(K - 1\right)} \to \max$$

Entropy: $$\left[PE = \frac{\sum_{q=1}^{Q}\sum_{k=1}^{K} \ln\left(u_{qk}\right)}{Q},\ PE \in \left[0, \ln Kk\right] \right]$$

Step 9. Calculation of the external balance criterion: $$ECB = \sqrt{\frac{\left(IC_A - IC_B\right)^2}{\left(IC_A + IC_B\right)^2}} \to opt$$

Step 10. If the value of the balance criterion reaches the optimum, then:

Step 11 Fixes the resulting clustering,

otherwise the number of clusters increases by 1 and steps 5–9 are repeated

Step 12. Determining the optimal number of clusters k_{opt}.

Step 13. Clustering data (the set Ω of objects under study), fixing the clusters.
Step 14. Validation of the results of clustering.
Step 15. Visualize the results of clustering.
Step 13. The End

4.5 Inductive Fuzzy C-Means Algorithm

The method of fuzzy clustering of c-means (or fuzzy clustering, soft k-means, c-means) allows you to split the existing set of elements with cardinality into a given number of fuzzy sets. The fuzzy clustering method of c-means can be considered as an improved method of k-means, in which for each element from the considered set the degree of its belonging (or responsibility) to each of the clusters is calculated. The algorithm was developed by J.C. Dunn in 1973 [21] and improved by J.C. Bezdek in 1981 [22].

1: Inductive algorithm (Ω , k)
2: normalization_canserSubtypes (Ω)
3: $\Omega := \{x_{ij}\}; \; i := \overline{1,n}; \; j := \overline{1,m}$
4: $\Omega^A := \{x_{ij}^A\}; \; \Omega^B := \{x_{ij}^B\}; \; i := \overline{1,n_A = n_B}; \; n_A + n_B = n; \; j := \overline{1,m}$
5: $k^A = k_{\min}^A, \; k^B = k_{\min}^B$
6: do
7: kmeans_clustering(Ω^A, k^A), kmeans_clustering(Ω^B, k^B)
8: $SWC^A := index\_silhouette\left(k^A\right), \; SWC^B := index\_silhouette\left(k^B\right)$
9: $DI^A := index\_Dunn\left(k^A\right), \; DI^B := index\_Dunn\left(k^B\right)$
10: $CH^A := index\_Calinski\_Harabasz\left(k^A\right), \; CH^B := index\_Calinski\_Harabasz\left(k^B\right)$
11: $PE^A := entropy\left(k^A\right), \; PE^B := entropy\left(k^B\right)$
12: $k^A = k^A + 1, \; k^B = k^B + 1$
13: $ECB := \sqrt{\left(IC^A - IC^B\right)^2} \Big/ \sqrt{\left(IC^A + IC^B\right)^2}$
14: while($ECB \to opt$)
15: $k_{opt} := \text{NbClust}\left(k^A, k^B\right)$
16: kmeans_clustering(Ω, k_{opt})
17: validation_clustering
18: visualization_result
19: end

(a)

1: Inductive algorithm (Ω , k)
2: normalization_canserSubtypes (Ω)
3: $\Omega := \{x_{ij}\}$; $i := \overline{1,n}$; $j := \overline{1,m}$
4: $\Omega^A := \{x_{ij}^A\}$; $\Omega^B := \{x_{ij}^B\}$;
$i := \overline{1,n_A} = n_B$; $n_A + n_B = n$; $j := \overline{1,m}$
5: $k^A = k_{\min}^A$, $k^B = k_{\min}^B$
6: do
7: cmeans_clustering(Ω^A, k^A), cmeans_clustering(Ω^B, k^B)
8: $SWC^A := index\_silhouette(k^A)$, $SWC^B := index\_silhouette(k^B)$
9: $DI^A := index\_Dunn(k^A)$, $DI^B := index\_Dunn(k^B)$
10: $CH^A := index\_Calinski\_Harabasz(k^A)$, $CH^B := index\_Calinski\_Harabasz(k^B)$
11: $PE^A := entropy(k^A)$, $PE^B := entropy(k^B)$
12: $k^A = k^A + 1$, $k^B = k^B + 1$
13: $ECB := \sqrt{\frac{(IC^A - IC^B)^2}{(IC^A + IC^B)^2}}$
14: while($ECB \to opt$)
15: $k_{opt} := \text{NbClust}(k^A, k^B)$
16: cmeans_clustering(Ω, k_{opt})
17: validation_clustering
18: visualization_result
19: end

(b)

Fig. 2. Pseudocode of inductive algorithm k-means (a) c-means (b)

4.6 Clustering Quality Assessment

As criteria for the quality of clustering were used:

1. Silhouette [23]

$$SWC = \frac{1}{K}\sum_{i=1}^{K} S_{x_j} \to \max$$

where K is the number of clusters, S_{x_j} is the "best" membership of the element x_j in the cluster p.

The best partition is characterized by the maximum SWC, which is achieved when the distance inside the cluster is small and the distance between the elements of neighboring clusters is large.

2. Dunn Index [24]

Compares intercluster dissolution with cluster diameter. The higher the index value, the better the clustering.

$$DI(k) = \min_{i \in k}$$

3. Calinski Index - Harabasz [25]

$$QC_{CH} = \frac{QCB \cdot (N-K)}{QCW \cdot (K-1)} \to \max$$

where N is the number of objects, K is the number of clusters. The maximum index value corresponds to the optimal cluster structure.

4. Entropy [26]

$$\left[PE = \frac{\sum_{q=1}^{Q} \sum_{k=1}^{K} \ln(u_{qk})}{Q},\ PE \in [0, \ln Kk] \right]$$

Entropy is known as a numerical expression of the ordering of a system. The entropy of the partition reaches a minimum at the highest ordering in the system (in the case of a clear partition, the entropy is zero). That is, the greater the degree of belonging of an element to one cluster (and the less the degree of belonging to all other clusters), the lower the value of entropy and the more qualitatively the clustering is performed.

5 Experiments and Results

For the experiment, we used data from the CancerSubtypes package, which is designed to assist in identifying the validation of cancer subtypes based on arrays of genomic cancer data. The package is implemented in the R language and is available as a bioconductor package at http: // bioconductor. org / packages / CancerSubtypes /. Glioblastoma gene expression data set (GBM) downloaded from TCGA. This is a small data set with 1,500 genes and 100 cancer samples extracted from gene expression data.

The inductive clustering algorithm was used for a complete set of data (1500x100), the preprocessing of which was carried out in the form of normalization (median, Z-score) and for data whose preprocessing includes normalization and reduction of data sizes based on analysis of the main components (PCA). As a result, the number of genes was reduced to 44 components (44x100).

For the experiment, two clustering algorithms were used - k-means and c-means. The results are presented in Table 1.

Table 1. Table captions should be placed above the tables.

Algorithm	K-means				C-means			
Data	GeneExp with median normalization	GeneExp with zscore normalization	PCA with median normalization	PCA with zscore normalization	GeneExp with median normalization	GeneExp with zscore normalization	PCA with median normalization	PCA with zscore normalization
Size	1500x100	1500x100	44x100	44x100	1500x100	1500x100	44x100	44x100
Clusters	3	3	3	3	3	3	3	2
Clusters size	46 54 46 89	45 25 70 478	42 1 1	13 16 15	61 84 53 429	61 84 15 467	18 22 4	20 1 23
Average Silhouette	0.06043241	0.07155017	0.00344284	0.00979075	0.05936179	0.07011326	0.000994238	0.007407239
Index Dunn	0.3385516	0.3620989	0.7851683	0.7982813	0.3385516	0.3620989	0.7524122	0.7945162
Entropy	1.090765	1.093573	0.2164141	1.094951	1.084931	1.084126	0.9302168	0.6890092

The clustering results were visualized using the Silhiuette method. This method allows to interpret and verify the data consistency within clusters. The technique provides a concise graphical representation. The Silhiuette value is a measure of how much the object resembles the one in its own cluster (cohesion) compared to other clusters (separation). The Silhiuette ranges from -1 to +1. A high positive value indicates that the object is in good agreement with its own cluster and is poorly aligned with neighbouring clusters. If most objects have a high positive value, then the clustering configuration is appropriate. If many points have a low or negative value, then too many or too few clusters can be in the clustering configuration. Figures 3 and 4 show a graphical representation of the clustering results of the inductive k-means and c-means algorithms.

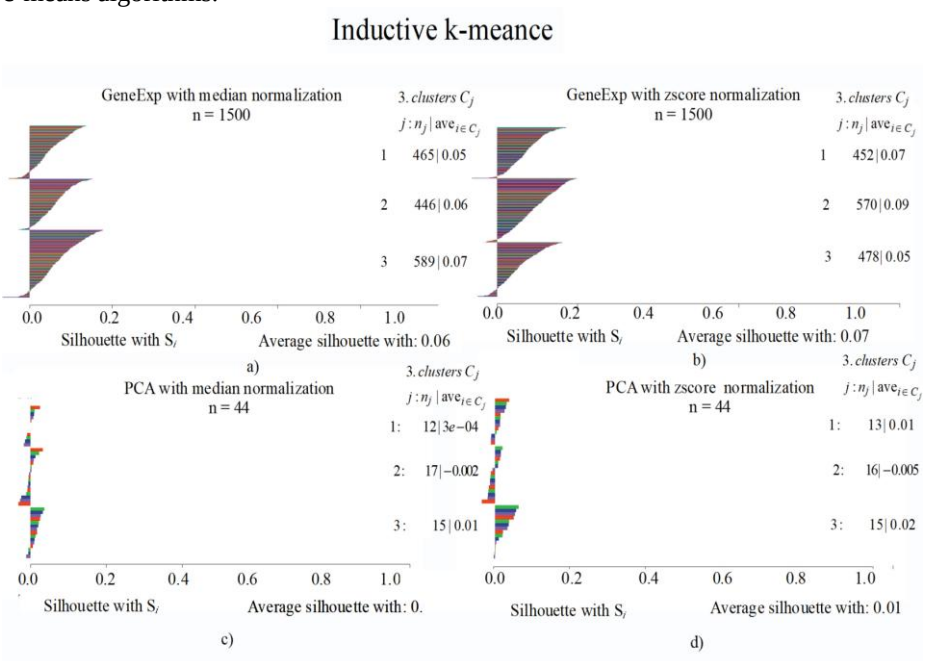

Fig. 3. The Silhiuette graphical representation when evaluating data clustering results using clustering by the inductive k-means algorithm after a median (a) and zcore (b) normalizing the data, as well as reducing this data using the PCA algorithm: median normalization + PCA (c) and zscore + PCA (d).

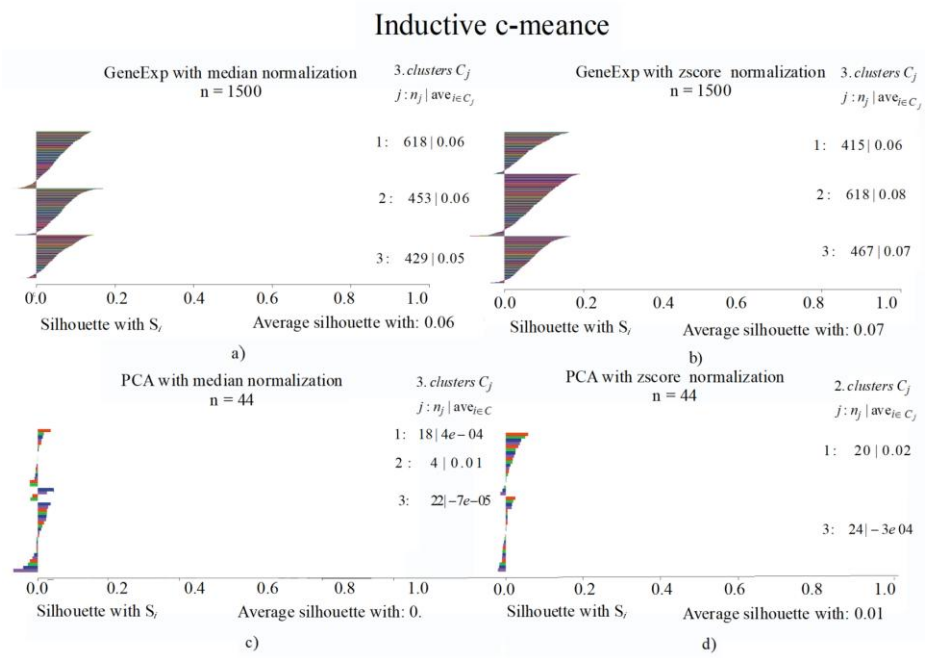

Fig. 4. The Silhiuette graphical representation when evaluating data clustering results using clustering by the inductive c-means algorithm after median (a) and zcore (b) normalizing the data, as well as reducing this data using the PCA algorithm: median normalization + PCA (c) and zscore + PCA (d).

6 Discussion

An analysis of the results allows us to conclude that the process of data pre-processing — normalization and size reduction using the principal component method — together with the inductive clustering algorithm improves the quality of clustering in terms of internal quality criteria.

The inductive k-means algorithm applied to GeneExp data that went through PCA pre-processing with median normalization gives the best result in terms of entropy, a satisfactory result on the Dunn index.

The best result for the average silhouette value is given by GeneExp with zscore normalization data with the inductive k-means algorithm.

The best result from the point of view of the Dunn index was obtained with PCA data with zscore normalization with the inductive k-means algorithm.

7 Conclusion

This paper presents the results of studies identifying the validation of cancer subtypes based on arrays of genomic cancer data. As an experiment, we used the GeneExp dataset obtained from the data from TCGA (The Cancer Genome Atlas) projects.

The source data matrix contained 1,500 genes and 100 cancer samples. At the first stage, we normalized the genes. In the second stage, the number of genes was changed to 44 components using the principal component analysis (PCA). Then we performed the inductive clustering algorithm and compared various clustering methods using internal clustering quality criteria as a criterion for evaluating the effectiveness of the corresponding clustering method.

An analysis of the processed data allows us to conclude that the proposed method is highly effective since its implementation can significantly reduce the set of components of cancer genomic data for subsequent processing.

References

1. Bleeker, F. E ., Molenaar, R. J., Leenstra, S.: Recent advances in the molecular understanding of glioblastoma. Journal of Neuro-Oncology. 108 (1): 11–27. PMC 3337398. PMID 22270850 doi: 10.1007 / s11060-011-0793-0 (2012).
2. Verhaak, R. G. et al.: Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. Cancer Cell. 17 (1), 98-110 (2010).
3. Network, T. C.: Corrigendum: Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature. 494 (7438), 506 (2013).
4. Frattini, V. et al.: The integrated landscape of driver genomic alterations in glioblastoma. Nat. Genet. 45 (10), 1141–1149 (2013).
5. The Cancer Genome Atlas homepage. NCI and the NHGRI. Retrieved 2009-04-28. http://cancergenome.nih.gov/
6. Cha, Y.J., Park, S.M., You, R., Kim, H., Yoon, D.K.: Microstructure arrays of DNA using topographic control. Nature Communications, 10(1), art. no. 2512 doi: 10.1038/s41467-019-10540-2(2019)
7. Lian, B., Hu, X., Shao, Z.-M.: Unveiling novel targets of paclitaxel resistance by single molecule long-read RNA sequencing in breast cancer. Scientific Reports, 9(1), art. no. 6032 doi: 10.1038/s41598-019-42184-z (2019)
8. Wang, Z., Gerstein, M., Snyder, M. :RNA-Seq: a revolutionary tool for transcriptomics. Nature Reviews. Genetics. 10 (1): 57–63. doi: 10.1038 / nrg2484. PMC 2949280. PMID 19015660 (2009).
9. Xu, T., Le, T. D., Liu, L., et al.: Identifying cancer subtypes from mirna-tf-mrna regulatory networks and expression data [J]. PloS one, 11 (4): e015279 (2016)
10. Omran, M., Engelbrecht, A., Salman, A.: An overview of clustering methods. Intell. Data Anal. 11(6): 583-605(2007)
11. Celebi, M. E ., Kingravi, H. A ., Vela, P. A.: A comparative study of efficient initialization methods for the k-means clustering algorithm. Expert Systems with Applications. 40 (1): 200-210. arXiv: 1209.1960(2013)
12. Madala, H. R.:Inductive Learning Algorithms for Complex Systems Modeling. CRC Press, 365 p. (1994)

13. Babichev, S., Taif, M. , Lytvynenko, V.: Estimation of the inductive model of objects clustering stability based on the k-means algorithm for different levels of data noise. Radio electronics, computer science, management. Zaporozhye: NAS of Ukraine, no. 4. pp.54-60 (2016)
14. Stepashko, V.S.: Elements of Inductive Modeling Theory - State and Prospects of Informatics Development in Ukraine: Monographic arts. Scientific Thought, pp. 471–486 (2010)
15. Osypenko, V. V.: The Methodology of Inductive System Analysis as a Tool of Engineering Researches Analytical Planning. Ann. Warsaw Univ. Life Sci. SGGW. no. 58. pp. 67–71 (2011)
16. Sarycheva, L.V.: Objective cluster analysis of the data on the basis of the Group Method of Data Handling. Problem of Management and Informatics, no. 2, pp. 86-104. [In Russian] (2008)
17. Steinhaus, H. :Sur la division des corp materiels en parties. Bull. Acad Polon Sci 1.804 (1956)
18. Lloyd, S. P. :Least square quantization in PCM. Bell Telephone Laboratories Paper. Published in journal much later: Lloyd, SP: Least squares quantization in PCM. IEEE Trans. Inform. Theor. (1957/1982).
19. MacQueen, J.: Some methods for classification and analysis of multivariate observations. Proceedings of the fifth Berkeley symposium on mathematical statistics and probability. vol. 1, no. 14(1967)
20. Melnik, M.: Fundamentals of applied statistics. Moscow: Energoatomizdat, 416 p. (1983)
21. Dunn, J.C.: A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters . Journal of Cybernetics. 17 09 (t. 3, no. 3), pp. 32–57. ISSN 0022-0280. - DOI: 10.1080 / 01969727308546046 (1973)
22. Bezdek, J.C.: Pattern Recognition with Fuzzy Objective Function Algorithms. ISBN 0-306-40671-3 (1981)
23. Kaufman, L., Rousseeuw, P.: Finding Groups in Data. An Introduction to Cluster Analysis. Wiley (2005)
24. Bezdek, J.C., Dunn, J.C.: Optimal fuzzy partitions: A heuristic for estimating the parameters in a mixture of normal dustrubutions. IEEE Transactions on Computers, 835–838(1975)
25. Calinski, R.B., Harabasz, J.: A dendrite method for cluster analysis // Comm. in Statistics, 3: 1.27(1974)
26. Sripada, S., Rao, G.: Comparison of purity and entropy of k-means clustering and fuzzy with means clustering, Indian journal of computer science and engineering, vol 2, no.3 ISSN: 0976-5166 (2011)