DOI:10.34918/85113

Mateusz KANIA[1,*], Andrzej POLAŃSKI[1]

Chapter 13. UNSUPERVISED CLUSTERING FOR DETECTION OF GENE EXPRESSION PATTERNS IN HUMAN CANCERS

13.1. Introduction

Cancer evolution is a complex dynamical process of uncontrolled growth of tissues/cells with dysregulated signalling, metabolism, and replication mechanisms. It is caused by somatic alterations/mutations of DNA, which can cumulate during mitotic replication of cells. Some studies report that somatic mutation might influence gene expression levels. Our study provides reliable, highly statistically significant support for gene expression pattern occurrence in human cancer. Our analysis is based on the Cancer Genome Atlas (TCGA) (Genomic Data Commons) database.

Our data is focused on gene expression levels in cancer since the matter of gene expression itself is still an extensively researched topic. We are studying the hypothesis that gene expression profiles would allow us to distinguish between different types of cancer. Another goal is to decide which unsupervised clustering algorithms will perform the best in the given task. The criterion performance in various metrics is calculated using clustering results and the ground truth.

13.2. Methods

For the experiment, we used the data from cBioportal, a portal for cancer genomics data. It is related to TCGA (The Cancer Genomic Atlas) being an interactive resource for the exploration of multivariate cancer genomic data. Moreover, cBioPortal

[1] Department of Computer Graphics, Vision and Digital Systems, Silesian University of Technology, Gliwice, Poland.
* Corresponding author: mateusz.kania@polsl.pl.

provides open access to molecular profiles and clinical attributes of different cancer genomic studies.

The resources of cBioPortal contain but are not limited to DNA methylation data, mRNA and microRNA expression or phosphoprotein level data (RPPA). We have used mRNA (messenger RNA) expression data for our analysis. mRNAs are the product of DNA transcription. The central role of messenger RNA is to function as a template for translation. During this process, mRNA sequences are first translated to amino acids, which then build functional proteins. Increased or decreased mRNA levels might be related to various diseases, including cancer. We wanted to determine if this kind of data contains enough information to distinguish between different types of cancer. The data we have used consisted of the median expression level of RNA sequencing data. We parsed the data to the format presented in Table 2. In rows, we gathered different cancer types mentioned in Table 1.

Table 1

Types of cancers

Cancer type
Stomach adenocarcinoma
Glioblastoma multiforme
Lung squamous cell carcinoma
Lung adenocarcinoma
Breast invasive carcinoma
Ovarian serous cystadenocarcinoma
Brain lower grade glioma
Thyroid carcinoma
Prostate adenocarcinoma
Pancreatic adenocarcinoma

Each row contains a subject, and each column describes the case using gene expression information. Next, we mixed gene expression information from all selected cancer types in possible combinations without repetitions. We created 50 mixtures of 2, 3, 4, 5 and 6 components. Each mixture was created ten times, containing different types of cancer. In addition, each set consisted of almost 20 000 features, and the number of observations ranged between 500-4000. Since the data is categorical and data points belong to real numbers, we assumed that this type of data might be described by a mixture of gaussian distributions.

Table 2

Input data – simplified example

	Gene 1	Gene 1	Gene 2
Cancer 1	15.4	15.4	5.2
Cancer 1	11.67	10.6	2.57
Cancer 2	18.67	12.64	1.8

13.2.1. Pipeline

The very first step in the analysis was data preprocessing. The data was parsed to fixed matrix format N × M where N indicates the number of patients and M number of genes, our variables of interest. Since we had to analyze more than 20 000 features, we used the decomposition method based on a Gaussian mixture of variances. According to the model, the features with the highest variance were left in the dataset, and the rest was treated as noise. In our comparison, we used four unsupervised algorithms: k-means, hierarchical agglomerative clustering, fuzzy c-means, and Gaussian EM. The listed algorithms are based on distance metrics, while the last uses a mixture of Gaussian distributions. To measure algorithms efficiency, we combined a few approaches. One was to prepare a binomial test for each algorithm. The other one was calculating Adjusted Rand Index, Simple Matching Coefficient, its weighted version and Averaged Jaccard Index. We applied the Hungarian algorithm, for the SMC index, which allowed us to assign classification results to their respective clusters.

13.2.2. Hierarchical clustering

Hierarchical clustering (HC) is called such because of the way it creates the clusters. HC results are a series of partitions with a visible hierarchy resembling tree branches. In the analogy, each branch is a cluster. The bigger branch consists of many smaller branches; at the end, it becomes a trunk, that incorporates all the data. However, there are two ways to cluster the data. We can start from a single point up to the whole data set or the complete data and successively build clusters up to one point. It is called agglomerative and divisive clustering. In our analysis, we are using the agglomerative method [3].

Hierarchical Agglomerative Clustering (HAC) is the most popular way to cluster the data. We first must decide upon the distance and linkage methods we will use during the analysis. The distance method explains how the distance or similarity between points will be measured. Some commonly known measures are Euclidean, Manhattan, Minkowski and others. The second choice is the linkage method. It determines how data points will be grouped in consecutive clusters. The few examples here will be single, complete or average linkage. The hierarchical clustering model allows us to choose any number of clusters without the need to repeat the calculations. It is a unique attribute of HC, not present in other unsupervised algorithms. Initially, we need to determine the metric that describes a relation between the data points. To do that, we can use similarity (e.g., Jaccard index) or distance (e.g., Euclidean) measure. The choice depends on scientific questions and the data itself. Next is the choice of linkage method. It will determine the way of how the data will be clustered together. It has a heavy impact on the results.

Last but not least thing to do is choose the number of clusters. We can do it in two different ways. The first one is to choose the number of groups exactly. Another way is to cut the branches at a specific tree height. The tree's height depends on the largest distance or similarity between two clusters in the data. We can also base our choice of clusters on this metric [2].

$$d_E(X,Y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots (x_n - y_n)^2} = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

$$d_E(X,Y) = d_E(X,Y)$$

where:

d_E – Euclidean distance

X, Y – probability distributions

x_i, y_i – realizations of X and Y

As for the linkage method, we used Ward's method that is based on minimizing error sum of squares (ESS).

$$\text{ESS}\ (X) = \sum_{i=1}^{N_X} \left| x_i - \frac{1}{N_X} \sum_{j=1}^{N_X} x_j \right|^2$$

$$d(X,Y) = ESS(XY) - [ESS(X) + ESS(Y)]$$

13.2.3. k-means

K-means algorithm is an iterative, distance-based algorithm. It is easy on resources, quick and computationally effective. Because of the low usage of memory, k-means is suitable for clustering huge data sets. It is a significant advantage over hierarchical clustering methods. Apart from their informative tree structure, they are computationally heavy.

Moreover, a k-means algorithm might be used to initialize other algorithms, for example, those based on Expectation-Maximization. From the mathematical perspective, k-means is similar to the normal mixture model. Estimation of parameters is done by the maximum likelihood method. The primary idea behind the k-means is that observations are gathered around artificially introduced centres, called centroids. Centroid can be treated as a mean generalization, a geometric centre of a convex object. In general, the distance between centres and observations should be minimal. Data points closest to the particular centre are part of its cluster [3]. The initial number of centroids is equivalent to the number of clusters k in the data. The number of initial groups is required to start the algorithm. There are different ways to choose the number of clusters beforehand, but we can also use expert knowledge or assumptions. The algorithm stops in a few cases. The most desirable one is the occurrence of convergence. For example, the creation of clusters with the highest similarity of points within a given cluster and the lowest between different ones. In the commonly used Hartigan-Wang algorithm, the stop criterion is based on minimizing the total sum of variance within clusters (WCSS). It is given by the formula [5].

$$WCSS = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \|x_{ij} - c_i\|^2$$

There are a few k-means algorithms: Lloyd, Forgy, MacQueen and the one already mentioned, Hartigan-Wong. The last one is the default k-means algorithm in the R software, used in the study.

13.2.4. Multivariable Gaussian mixture clustering (GaussEM)

Fitting the multivariable Gaussian mixture model to data can be done using the Expectation-Maximization algorithm was (Dempster, Laird and Rubin [1]). It is commonly used in a situation when the observations can be viewed as incomplete.

Some examples of cases when it is used are: missing data, truncated distributions, censored or grouped data [7]. The usual requirement to start EM for Gaussian mixture is to provide a k number of clusters. Knowing the number of subgroups in the data, we can initialize parameters in the next step. Initialization During initialization, we need to create a first guess of the parameters. In the case of Gaussian mixture, we need to initialize mixing proportions (α), mean (p), and variance σ 2 for each mixture component k ∈ {1..K}. Mixing proportions indicates how much of the mixture space belongs to K. Depending on the number of K, we need to provide equal number of α_K.

$$\alpha_1 \quad \alpha_2 \quad \alpha_3 \quad \cdots \quad \alpha_K$$

Assume that we have a mixture where k = 3. In that case, we need to create three α parameters. We can use uniform distribution $\alpha_K \sim U(0.1, 1)$ to obtain initial alphas. We advise to keeping the interval within [0.1, 1] because shallow values might cause over dominance of larger α during the estimation step. After choosing values from a uniform distribution, they should be standardized.

$$\hat{\alpha}_K = \frac{\alpha_1 + \alpha_2 + \cdots + \alpha_k}{\sum_{k=1}^{K} \alpha_k} \text{ and } \sum_{k=1}^{K} \hat{\alpha}_k = 1$$

The Gaussian probability density function is expressed as:

$$f_{k,m}(x) = \frac{1}{\sqrt{2\pi}\sigma_{k,m}} \exp\left[-\frac{\left(x - \mu_{k,m}\right)^2}{2\sigma_{k,m}^2}\right]$$

The E-step utilizes Bayes Theorem. Likelihood of data, given model is multiplied by prior value, alpha. Alpha is treated as a mixing proportion value.

$$p(k \mid x_n^1, \ldots, x_n^M, p^{\text{old}}) = \frac{\alpha_k^{\text{old}} \prod_{m=1}^{M} f_{k,m}(x_n^m, p^{\text{old}})}{\sum_{\kappa=1}^{K} \alpha_\kappa^{\text{old}} \prod_{m=1}^{M} f_{\kappa,m}(x_n^m, p^{\text{old}})}$$

p is equal to the vector of required parameters, $p = [\mu_{1,1}, \ldots, \mu_{1,M}, \ldots, \mu_{K,1}, \ldots, \mu_{K,M}, \sigma_{1,1}, \ldots, \sigma_{1,M}, \ldots, \sigma_{K,1}, \ldots, \sigma_{K,M}]$

The M-step is used to update parameters α, μ and σ^2, according to the presented equations.

$$\alpha_k^{\text{new}} = \frac{\sum_{n=1}^{N} p\left(k \mid x_n^1, \ldots, x_n^M, p^{\text{old}}\right)}{N}$$

$$\mu_{k,m}^{\text{new}} = \frac{\sum_{n=1}^{N} x_n^m p\left(k \mid x_n^1, \ldots, x_n^M, p^{\text{old}}\right)}{\sum_{n=1}^{N} p(k \mid x_n^1, \ldots, x_n^M, p^{\text{old}})}, \; k = 1,2, \ldots, K$$

$$\left(\sigma_{k,m}^{\text{new}}\right)^2 = \frac{\sum_{n=1}^{N} \left(x_n - \mu_k^{\text{new}}\right)^2 p\left(k \mid x_n^1, \ldots, x_n^M, p^{\text{old}}\right)}{\sum_{n=1}^{N} p\left(k \mid x_n^1, \ldots, x_n^M, p^{\text{old}}\right)} k = 1,2,\ldots,K, m = 1,\ldots,M$$

The algorithm finishes its iterations when the absolute difference between old and new parameters is less than 1e-6.

13.2.5. Adjusted Rand Index

To determine efficiency of algorithms, we compared them using the Adjusted Rand Index (ARI) index. We have implemented the version that was proposed in [13].

$$ARI = \frac{\sum_{ij}\binom{n_{ij}}{2} - \left[\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}\right] / \binom{n}{2}}{\frac{1}{2}\left[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}\right] - \left[\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}\right] / \binom{n}{2}}$$

where:

where:	Y_1	Y_2	$\cdots$	Y_s	sums
X_1	n_{11}	n_{12}	$\cdots$	n_{1s}	a_1
X_2	n_{21}	n_{22}	$\cdots$	n_{2s}	a_2
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
X_r	n_{r1}	n_{r2}	$\cdots$	n_{rs}	a_r
sums	b_1	b_2	$\cdots$	b_s	

13.2.6. Simple Matching Coefficient and its weighted version

To calculate Simple Matching Coefficient (SMC) and Weighted Simple Matching Coefficient (WSMC), we have used the Hungarian algorithm. However, instead of finding a minimal value for each row/column, we were looking for a maximal value. In this way, the maximal value was considered a true positive. The Simple Matching Coefficient is a straightforward metric. We divide the sum of true positives by the number of all values. The downside of this solution is that it does not impose any weights [10]. The green colour in Fig. 1 indicates matching groups.

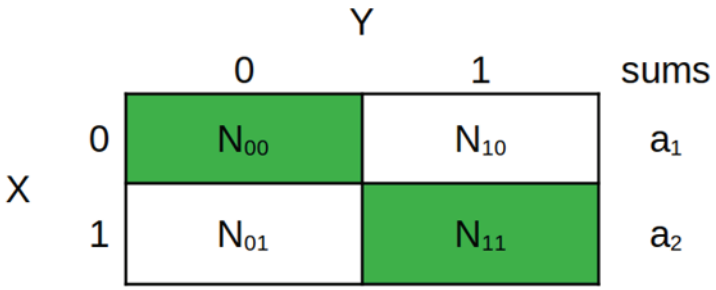

Fig. 1. Cross-table of clustering results
Rys. 1. Tabela krzyżowa wyników grupowania

Simple Matching Coefficient equation:

$$\mathrm{SMC} = \frac{N_{00} + N_{11}}{N_{00} + N_{11} + N_{01} + N_{10}}$$

To mitigate the problem with unequal groups, we can use the WSMC metric that addresses this issue:

$$\mathrm{WSMC} = \frac{N_{00}}{a_1} + \frac{N_{11}}{a_2}$$

13.2.7. Averaged Jaccard index

The Jaccard index is a popular metric [4]. Here we are using its simple variation to take two and more clusters.

$$\mathrm{J} = \left(\frac{N_{00}}{N_{11} + N_{01} + N_{10}} + \frac{N_{11}}{N_{00} + N_{01} + N_{10}}\right)/2$$

With the increased number of clusters, we have more terms in brackets and the denominator changes accordingly.

13.2.8. Binomial test

We used a binomial test to test hypotheses about correctly classifying cancer. It was possible since the classification result might have two outcomes: success or failure. The probability p was scaled accordingly to the number of clusters with $p = 1/k$.

13.3. Results and discussion

To show the efficacy of algorithms from diverse perspectives, we present results using various index values.

We start with the classical one, the Adjusted Rand Index.

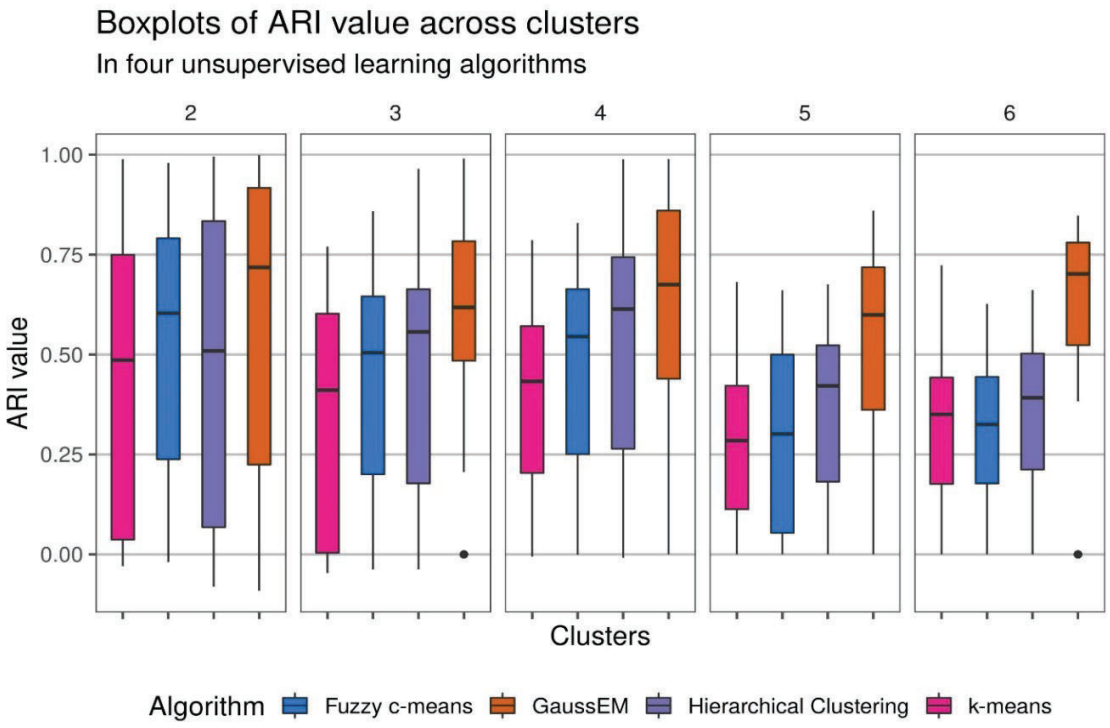

Fig. 2. Performance of four algorithms in various number of clusters
Rys. 2. Wydajność czterech algorytmów przy zmiennej ilości grup

In Fig. 2, in the case of HC, k-means and fuzzy c-means, we observe that the ARI index decreases with the increased number of groups. This trend is barely observed in the case of the GaussEM algorithm. Here, the median value oscillates between 0.74 and 0.60.

We can use violin plots to make the results more compact (Fig. 3). In addition to the values, violin plots show their density. The first three algorithms are denser. In the case of Gauss EM, the values are more concentrated in the upper parts of value indexes.

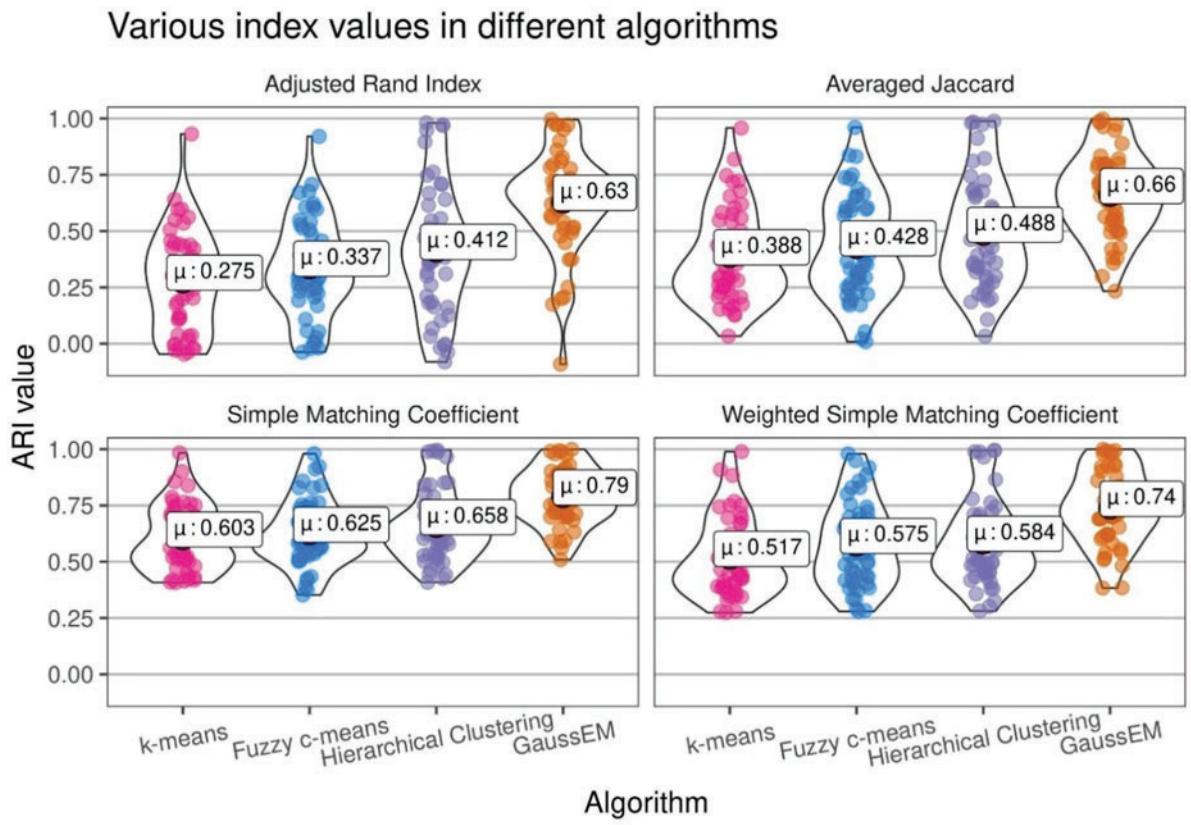

Fig. 3. Violin plots showing different quality measures
Rys. 3. Wykresy skrzypcowe pokazujące różne miary jakości

Fig. 4 shows results from the binomial test. The p-values are scaled to the power of 1/60 to make the results visible. We can see that majority of the values are below the p = 0.05. What is worth noticing is that even after raising results to the power of 1/60, the median value of GaussEM remains close to zero.

We compared the performance of four different unsupervised clustering methods that were based on distance metrics and maximum likelihood. For the comparison, we used the gene expression data from the TCGA portal. To assess the efficacy of the algorithms, we used various metrics, like Adjusted Rand Index, Jaccard, Simple Matching Coefficient, Weighted Simple Matching Coefficient, and binomial test. All of the compared algorithms showed statistically significant results. HAC performed as second best. Although computationally heavy, it is not very useful for big data in its original form.

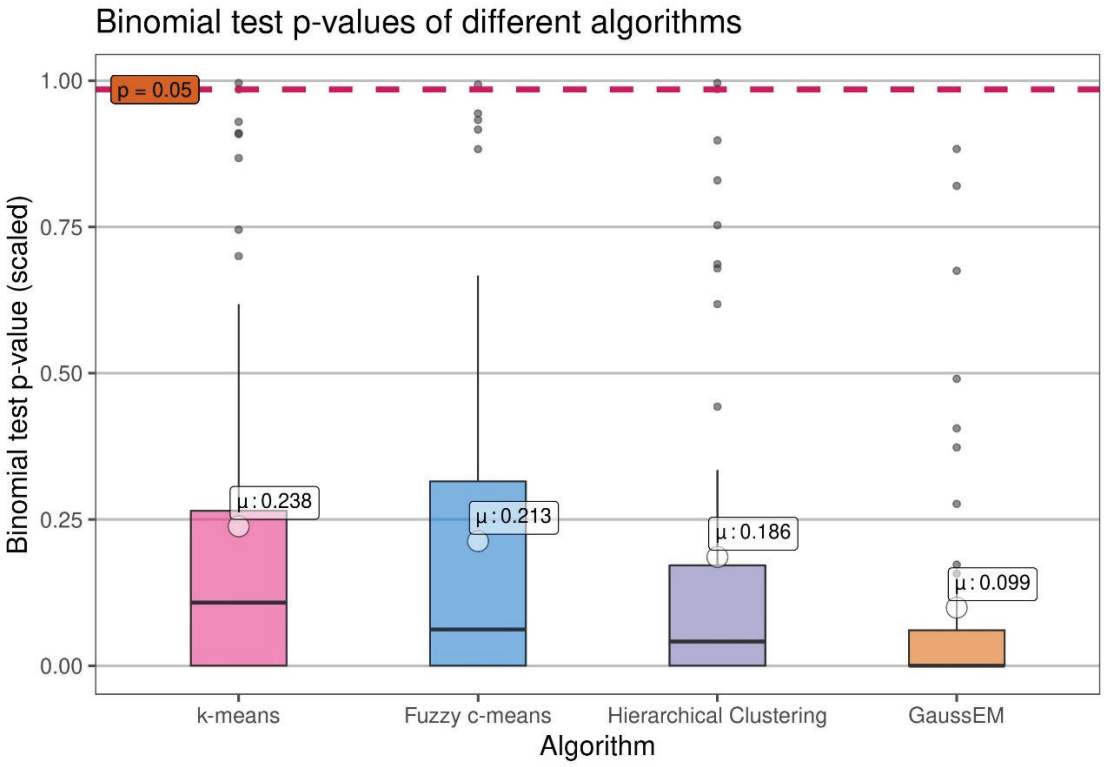

Fig. 4. Scores from the binomial test, raised to the power of 1/60
Rys. 4. Wyniki testu dwumianowego, podniesione do potęgi 1/60

Fuzzy c-means, a bridge between distance and maximum likelihood-based methods, performed better than k-means. Perhaps results might be improved using different parameters, e.g. fuzzyfied value. K-means scored the lowest both in the binomial test and other metrics. Although its simplicity, and low computational power, the requirement makes it still valuable, whether it is data exploration or starting point for the Expectation-Maximization methods. Finally, the multivariable Gaussian mixture with Expectation-Maximization algorithm obtained the highest score in all the presented metrics. It indicates that GaussEM is a good approach to finding gene expression patterns in human cancers.

Acknowledgment

This work is supported by the European Social Fund grant no. POWR.03.02.00-00-I029.

Bibliography

1. Arthur P. Dempster, Nan M. Laird, Donald B. Rubin: Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, 1977.
2. Jiarui Ding, Melissa K. McConechy, Hugo M. Horlings, Gavin Ha, Fong Chun Chan, Tyler Funnell, Sarah C Mullaly, Jueri Reimand, Ali Bashashati, Gary D. Bader, et al.: Systematic analysis of somatic mutations impacting gene expression in 12 tumour types. *Nature communications*, 6(1):1–13, 2015.
3. Paolo Giordani, Maria Brigida Ferraro, Francesca Martella: An Introduction to Clustering with R. *Springer*, 2020.
4. Lieve Hamers et al.: Similarity measures in scientometric research: The jaccard index versus salton's cosine formula. *Information Processing and Management*, 25(3):315–18, 1989.
5. Peilin Jia, Zhongming Zhao: Impacts of somatic mutations on gene expression: an association perspective. *Briefings in bioinformatics*, 18(3):413–425, 2017.
6. Leonard Kaufman, Peter J. Rousseeuw: Finding groups in data: an introduction to cluster analysis, volume 344. *John Wiley & Sons*, 2009.
7. Geofrey J. McLachlan, Thriyambakam Krishnan: The EM algorithm and extensions, volume 382. *John Wiley & Sons*, 2007.
8. Laurence Morissette, Sylvain Chartier: The k-means clustering technique: General considerations and implementation in mathematica. *Tutorials in Quantitative Methods for Psychology*, 9(1):15–24, 2013.
9. Joanna Polańska, Piotr Widłak, Joanna Rzeszowska-Wolny, Marek Kimmel, Andrzej Polański: Gaussian mixture decomposition of time-course DNA microarray data. In Mathematical Modeling of Biological Systems, Volume I, pages 351–359. *Springer*, 2007.
10. Jorge M. Santos, Mark Embrechts: On the use of the adjusted rand index as a metric for evaluating supervised classification. In International conference on artificial neural networks, pages 175–184. *Springer*, 2009.
11. Robert R. Sokal: A statistical method for evaluating systematic relationships. Univ. Kansas, *Sci. Bull*., 38:1409–1438, 1958.
12. Christopher Wilks, Melissa S. Cline, Erich Weiler, Mark Diehkans, Brian Craft, Christy Martin, Daniel Murphy, Howdy Pierce, John Black, Donavan Nelson, et al.: The cancer genomics hub (cghub): overcoming cancer through the power of torrential data. *Database*, 2014.
13. Lawrence Hubert and Phipps Arabie, Comparing partitions. *Journal of classification 2*(1):193–218, 1985.

UNSUPERVISED CLUSTERING FOR DETECTION OF GENE EXPRESSION PATTERNS IN HUMAN CANCERS

Abstract

In this study, we compare four unsupervised algorithms in the gene expression data of different human cancers. We based our analysis on openly available data from Cancer Genome Atlas (TCGA) (Genomic Data Commons) database. We tested two properties. The first is if there is a clear pattern in the gene expression data. The other was to select the algorithm which performs the best. Our results suggest that an expression pattern exists in different types of human cancer. As for the algorithm, the EM algorithm based on multivariate Gaussian mixtures showed the most promising performance.

Keywords: unsupervised clustering, GMM, k-means, HC, cancer, gene expression, EM