Clustering for Gene Expression Analysis

Olga Georgieva

Sofia University „St. Kliment Ohridski", 5 James Bourchier blvd., 1164 Sofia, Bulgaria

Abstract

Since several decades intensive research for revealing and understanding gene expression and consequent their role in organisms live are research focus of scientists in different areas. A part of these investigations is the analysis of gene expression data. The present paper extends and by that improves an iterative clustering procedure for detection genes that are differentially expressed. The presented analytical review of the clustering algorithms applied to the gene expression analysis serves as a good prerequisite for this aim. Additional investigation is presented and conclusion for appropriate choice of the procedure parameters is summarized.

Keywords

Gene expression analysis, Differential gene expression, Clustering analysis

1 Introduction

Since several decades intensive research for revealing and understanding gene expression and consequent their role in organisms live are research focus of scientists in different areas. A part of these investigations is the analysis of gene expression data. This is a numerical table of values of expression levels of thousands of genes observed for (usually large amounts) of samples [1].

The expression data tables are obtained through well-established technologies, which have been significantly improved in recent years. Thus, the contemporary technologies named Next Generation Sequencing provide large volumes of DNA-seq and RNA-seq data. The obtained gene expression data by them are more precise having higher resolution than the older technologies of microarray. It significantly increases the opportunities for effective research to gather knowledge for existing dependencies of the genes' activity. Despite the technology used to obtain gene expression, it results in gene expression table that contains up to several hundred thousand numbers, where each row corresponds to one particular gene and each column to a sample. The number of genes in such table comprises a part or whole genome of a given organism. This huge amount of data needs to be processed applying specific data mining algorithms to reveal new information about unknown biological dependences. Typical information that could be revealed by processing these tables is summarized in the following general tasks [1,5,6]:

- To understand how genome sequences specify the forms and functions in the organisms and how remarkable diversity of life appears on the Earth.
- To understand the mechanisms and evolution of the organism's response to environment changes.
- To find groups of genes that act equally and by that to identify distinct subsets of genes that are unknown in their biological response.

Specific solutions of the above general tasks could be mentioned as the aim to identify genes with expression levels that reflect biological processes of certain interest as for instance cancer. One can identify a specific type of cancer in respect to group of genes responsible for it. Thus, classification

Education and Research in the Information Society, October 13-14, 2022, Plovdiv, Bulgaria
EMAIL: ogeorgieva@fmi.uni-sofia.bg
ORCID: 0000-0001-5376-6729

obtained from gene expression analysis can be used for more precise tumor diagnostic and its effective treatment [1].

The posed general tasks need to discover groups of genes could be solved in a supervised manner if preliminary knowledge about these groups exists. However, in most cases it is searched for unknown data partitioning. In this case unsupervised methods of cluster analysis are appropriate tool [3,4]. Due to the vast amount of data in the gene expression matrix it is helpful to process data by subdividing the genes into a smaller number of categories and then analyzing the obtained groups [2]. It is more interesting to cluster genes that show similar expression patterns across a number of samples, rather than clustering the samples themselves.

A specific task of gene expression research is the aim for defining differentially expressed genes. These are genes that act differently in equal strains, for which many samples of each strain have been collected. In this way the genes responsible for certain disease states can be discovered or it is possible to find differences between two species strains. In searching an appropriate solution several methods for RNA-seq data analysis that identify different genes according to their expression levels basically on statistical data analysis have been explored [7,8,9]. However, there is no good agreement among the applied methods as beside the commonly identified genes each method detects additional genes that are not identified by the others. Furthermore, the additional genes are large number among the distinct methods. By taking the advantage of the unsupervised data analysis methods an iterative clustering procedure to deal with this task was recently introduced and compared with several statistical methods [10]. The difficulty of the selection problem due to the large number of indistinguishable genes is solved by iterative solutions. The procedure depends on several parameters, which values are subject of a preliminary choice.

The present paper aims to extend and by that to improve the iterative clustering procedure for differential gene expression detection. The analytical review presented in the paper underlines the powerful ability of the clustering algorithms to deal with gene expression data. The review serves as a good prerequisite to solve the research aim. Based on it additional investigation is presented and conclusion for appropriate choice of the procedure parameters is summarized.

2 Clustering analysis for gene expression data analysis

Clustering is machine learning method that represents the collection of data in groups/clusters. The data are similar to each other within the cluster and different from data of other clusters. There is large diversity of algorithms that are able to deal with this problem. Diversity is governed by the need to incorporate solutions to two important questions. First, what similarity measure to involve to assess the data proximity and second, what procedure to use in order to find the data groups. The first one is the most important factor that defines the shape of the determined clusters. According to [3] “There is no single best criterion for obtaining a partition” due to the fact that “each clustering criterion imposes a certain structure on the data”.

Despite the large diversity of clustering methods, they could be grouped according to the implemented clustering technique. Three main groups should be underlined for gene expression analysis – hierarchical clustering, partitioning according a criterion for proximity and density-based solutions.

2.1 Hierarchical clustering for gene expression analysis

The most applicable clustering method for gene expression analysis is hierarchical clustering [4,5]. Each cluster is built by smaller clusters, forming a tree-shaped data structure or dendrogram. Two strategies for tree development are known. Agglomerative hierarchical clustering starts with the single-gene clusters and successively joins the closest clusters until all genes have been joined into the supercluster. The other strategy starts in opposite manner – all data are collected in a single cluster and further they are divided into smaller groups. The agglomerative hierarchical algorithm that joins pairs of clusters on the basis of their proximity, is the most widely used for gene expression analysis [5].

The important question is the cutting level of the dendrogram in order to obtain the right clusters. Another difficulty concerns how data are linked to form a cluster. A family of clustering methods is

known according to the implemented linkage function. Due to the existing indistinguishability of genes in their expression different initial linking and different linkage functions have to be explored in order to establish good clustering results.

The establishment of hierarchical clustering method as a widely used method for clustering gene expression data is owed to its visibility features. The clustering results are presented in a dendrogram. However, they are also given in a colored view way that has become a standard for visualization of the gene expression data.

2.2 Objective function clustering for gene expression analysis

The most used objective function clustering algorithm is *k*-means. It divides the data into a predetermined number of *k* clusters. The algorithm identifies the clusters according to their representatives – cluster centers. Data points are assigned to a cluster on the basis of the distances from the centroids. For instance, an application for gene expression analysis is demonstrated in [11].

A family of objective functions algorithms based on *k*-means have been proposed. The major question of their application is how many clusters actually exist and thus to initialize the algorithm. The answer of this question is not trivial. It is a subject of additional search through extended procedure. One possible approach is to initialize with *k* randomly chosen cluster centroids, and each gene is assigned to the cluster with the closest centroid. Other good strategy is seeding prototype centroids with the eigen vectors identified by PCA performed optimally for genes of yeast bacteria [12]. Subtractive clustering was successfully implemented to determine the optimal number of clusters of gene expression of soil bacterium data [13].

The fuzzy variant of the *k*-means algorithm – Fuzzy C-means, shows large advantage for gene expression analysis as it finds clusters that are overlapped. It is a powerful tool to reflect the real relationship between genes pointing distinct regulations and features of each gene's function [12,13].

The self organizing map (SOM) methods are underlined to join these groups of clustering methods. They find clusters, which are organized into a grid structure. However, the search procedure follows the same idea of proximity assessment of the input vector and lattice representation [1,2].

2.3 Density-based algorithms for gene expression analysis

The density-based clustering searching for dense areas in the data space. It is not necessarily to generate clusters explicitly, but instead to show the bunches of data that form cluster structure of the data set. These algorithms are good way to separate clusters from the noise. It allows for centralized description of irregularly shaped clusters in a data set with high dimension to identify outliers as data points with low cardinality [14,15]. Such proposed algorithm [15] implemented in gene-based clustering the dense and nondense areas of data are revealed, which explains complexes and patterns of the gene associations. The implementation shows better efficacy than DBSCAN, which is another density-based clustering algorithm. On other hand, DBSCAN clustering [16] has a simple scheme for clusters detection. It uses a matrix of pairwise distances between data and finds a number of outliers and core points. The key clustering parameters are the threshold radius *r* for neighborhood search and a minimum number of neighbors *Nmin* required to identify a core point.

It should be underlined that other methods different from the discussed above have been proposed in the recent years for gene expression analysis. Pattern-based clustering algorithms form clusters by objects, which attributes present difference of changes of the attributes values smaller than a threshold value. Other workable idea was to construct and apply clustering algorithm by iterative manner, which is a helpful strategy in case of vast amount of data.

3 Clustering for differential gene expression analysis

Differential gene expression problem is slightly different from the general task of gene expression analysis, which tries to separate genes in groups by their similarity. Instead, the differential expression analysis aims to find genes with significantly different expression than the others with the same behavior [7,9].

The difficulty of the task could be summarized as:

- Due to the large number of genes and lack of significance in separation by the estimated *p*-values no valuable conclusion about the gene separation could be done.
- Often the number of searched genes is (quite) smaller than the whole genes' amount. These genes could be treated more as outliers than a significant group.
- The genes do not have (near) constantly activity levels among different samples of a certain strain. This imposes to work with average values than to single sample data.

A recently proposed solution takes advantages from the density-based clustering approach dueto its ability to find outliers with the core clusters. The implemented algorithm separates genes densely scabbed around the equivalence area from the outliers that are away from this area. In fact, the outliers are data of the interested differentially expressed genes. Figure 1 illustrates this observation for DBSCAN clustering applied to the average values of the gene's activity of two mice strains. However, only 154 genes were discovered against several hundred found by statistical data analysis.

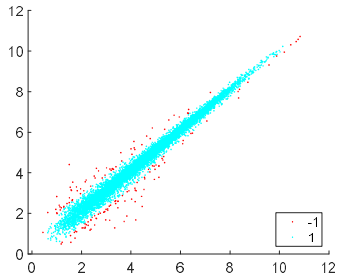

Figure 1: Results of DBSCAN, *r* = 0,4, *Nmin* = 20. Discovered clusters are enumerated in the picture legends: outliers "-1", core cluster "1" [10].

In order to overcome the problem of genes' undistinguishing as small number of differentially expressed genes could be found when clustering the entire data set, an iterative clustering scheme is proposed. By that the number of discovered genes significantly increases. The procedure motivation is detailly given in [10]. Its steps are briefly presented in the next section. In addition, the following text concerns further procedure improvement by more insight and analysis about the values of the procedure parameters.

3.1 Iterative clustering solution for differential gene expression analysis

The proposed clustering procedure [10] incorporates DBSCAN algorithm applied iteratively to genes data divided into batches. The outliers defined in each batch are added to form a common set of differently expressed genes.

3.2 Data preprocessing

Two main obstacles have to be tackled before clustering. First, logarithm transformation is needed in order to solve the scalability problem. Some of the genes' expressions are very large in respect to others. By that the distance estimation and then grouping will be distorted. Second, zero expression values are commonly seen in some genes, which were not active or their expression have been not measured. In order to ensure the logarithm calculation, filtering for removing genes with zero activity value is a requisite.

3.3 Data processing

Iterative implementation of DBSCAN requires setting the clustering parameters – the threshold radius *r* for neighborhood search, minimum number of neighbors *Nmin* and the number of genes' data that form the batch.

Again, due to the similarity in the genes' behavior as in case of the whole data set (Fig. 1), it could be expected that the data of each batch form a large compact group along the equivalence area. This area is rather oblong than spherical one. It suggests that clusters are not Euclidean. This observation requires further detail investigation into the selection of an appropriate proximity measure. In searching of proper distance measure different distances have to be applied and the results to be assessed in terms of separation abilities.

The minimum number *Nmin* is difficult to find in advance as it depends on the density of the data. Here we determine heuristically in terms to it increase the outlier's separation.

The number of genes that form a batch is other problem to be solved. The larger batch could make impossible to detect the outliers at all, whereas small batch could embarrass detection of clusters and thus the right distinguish between equivalent and differentially expressed genes.

Once the parameters are determined they are applied to each data batch according to the accepted iterative scheme of clustering.

3.4 Parameter analysis and application results

The iterative procedure of differently expressed genes' selection is applied for the set of samples of two mice strains – ten of strain C57BL/6J and eleven of strain DBA/2J, represented in data of 13932 genes having non-all-zero rows in the dataset [17]. As still there are zero expression values in some rows of the data, after the preprocessing stage the data set was reduced to 9196 genes.

First, we explore different distance metrics in the attempt to find the best method parameters (Table 1). For this the accumulated group of differently expressed genes marked by *ML* is compared with genes' groups separated by statistical data analysis of four statistical methods – *ttest*, *edgeR*, *limma*, *DESeq2*. The number of discovered genes by each statistical method of the filtered dataset provided by [7] is presented at the first (sub)column of the respective method column. The number of genes discovered by our procedure that are common for the respective statistical method is given at the second (sub)column. The last column of the table "*ML* all data" consists the total amount of differently expressed genes that are identified by the proposed iterative clustering method.

Several distance metrics have been investigated and those that produce good results are presented and discussed here. Both Euclidean distance as well as Minkowski distance with value of its parameter p=2 (or close to 2) give relatively good results according to the separated clusters. However, as they form spherical clusters that do not correspond to the data structure, it is a prerequisite to distort the separation result by mixing some outliers with the selected good clusters. The distances that produce oblong clusters are more valuable in respect to the real data structure. These are Mahalonobis, Minkowski with another parameter value than p = 2. The distances Cityblock and Chebishev are also investigated. By varying DBSCAN parameters different amounts of genes are discovered and the best results for each method are presented at the table. Notwithstanding the fact that Minkowski distance with small value p=0,5 presents best result according to commonly identified genes with respective statistical method, the preference is given to Mahalanobis distance clustering as well (both in bold). The oblong clusters through Mahalanobis distance determine the smallest number of 1905 selected genes.

Table 1: Number of differentially expressed genes selected by different cluster distances compared with results of different statistical methods

Method	***ttest***		***edgeR***		***limma***		***DESeq2***		***ML***
Distance	*ttest*	*ML*	*edgeR*	*ML*	*limma*	*ML*	*DESeq2*	*ML*	all data
Euclidean	71	71	915	647	736	537	982	648	2848
Mahalanobis	**71**	**71**	**915**	**738**	**736**	**611**	**982**	**735**	**1905**
Cityblock	71	70	915	417	736	365	982	407	994
Minkowski, p=2	71	68	915	286	736	266	982	283	614
Minkowski, p=0.5	**71**	**71**	**915**	**789**	**736**	**641**	**982**	**807**	**3475**
Chebychev, r=0,1, *Nmin*=5	71	71	915	667	736	548	982	665	2264

In searching the appropriate batch volume, we explore different volumes for clustering by best found cluster distance (Table 2). The results for batches consisting of 511 and 1022 genes with appropriate

settings of the rest two parameters – *r* and *Nmin,* are comparable. Certain preferences can be given to clustering at batch of 511 genes because of the smaller number of all separated genes (*ML* all data). On the other hand, batch of 1022 genes ensures larger number of differentially expressed genes.

Table 2: Number of differentially expressed genes selected by iterative procedure with different batch volumes size and applied Mahalonobis distance

	ttest		***edgeR***		***limma***		***DESeq2***		
Number of genes in a batch, algorithm parameters	*ttest*	*ML*	*edgeR*	*ML*	*limma*	*ML*	*DESeq2*	*ML*	***ML*** all data
511, *r*=0,2, *Nmin*=5	**71**	**71**	**915**	**738**	**736**	**611**	**982**	**735**	**1905**
511, *r*=0,3, *Nmin*=5	71	71	**915**	527	**736**	457	**982**	516	896
1022, *r*=0,2, *Nmin*=5	71	71	915	584	736	508	982	573	1058
1022, *r*=0,2, *Nmin*=10	**71**	**71**	**915**	**791**	**736**	**642**	**982**	**787**	**2050**
1022, *r*=0,3, *Nmin*=10	71	71	915	525	736	455	982	514	891

The advantage of the procedure is its result visibility abilities. For each batch the clustering result could be seen and assessed visually. Further assessment and interpretation according the biological meaning of the compact clusters found in the equivalent area and the respective outliers could be applied (Figure 2).

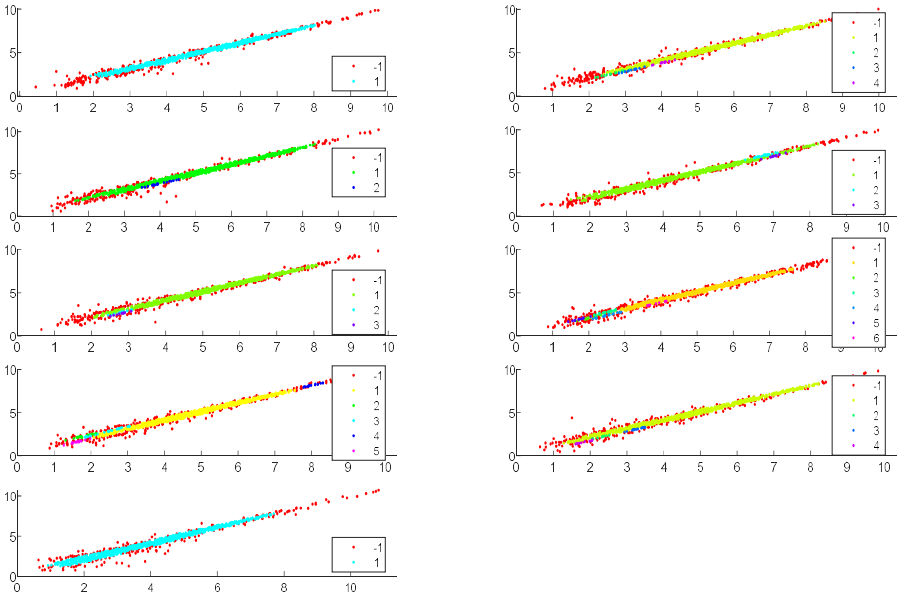

Figure 2: Results of iterative DBSCAN clustering by Mahalonobis distance measure, *r*=0,3, *Nmin*=10 and bach=1022 genes.

4 Conclusion

The paper improves an iterative clustering procedure for differential gene expression detection. For this aim, first an analytical review that underlines the powerful ability of the clustering algorithms to deal with gene expression data by revealing their advantages and disadvantages for solving gene expression data analysis tasks is presented. Second, the specific task for searching genes that are differentially expressed is considered. The task is solved by procedure taking the advantage both of density-based clustering and the iterative clustering.

The paper makes conclusions about the appropriate choice of the procedure parameters – proper distance measure and batch volume. The results obtained are compared with statistical data analysis applied to the investigated dataset of gene expression of two mice strains. The best distance measures are found that are Mahalonobis and Minkowski distances. Good results are obtained for batches of 511 and 1022 genes.

The iterative procedure is standalone applicable. In other cases, it could be used in combination with statistical methods in order to stick their search in a smaller number of genes.

5 Acknowledgements

This research work has been supported by GATE project, funded by the Horizon 2020 WIDESPREAD-2018-2020 TEAMING Phase 2 programme under grant agreement No.857155 and Operational Programme Science, by Operational Programme Science and Education for Smart Growth under Grant Agreement no. BG05M2OP001-1.003-0002-C01 and by the Science Fund of Sofia University under project no. 80-10-169/23.05.2022.

6 References

[1] E. Domany, Cluster Analysis of Gene Expression Data. Journal of Statistical Physics 110 (2003): 1117–1139.

[2] P. D'haeseleer, How does gene expression clustering work? Nat Biotechnol 23 (2005): 1499–1501.

[3] A.K. Jain, R.C.Dubes, Algorithms for Clustering Data. (Prentice Hall, Englewood Cliffs, New Jersey, 1988.

[4] D. Jiang, C. Tang, A. Zhang, Cluster analysis for gene expression data: a survey. IEEE Trans. Know. Data Eng. 16 (2004): 1370–1386.

[5] M.B. Eisen, P.T. Spellman, P.O. Brown, D. Botstein, Cluster analysis and display of genome-wide expression patterns. Proc. Natl. Acad. Sci. USA 95 (1998): 14863–14868.

[6] P.T.Spellman, Cluster analysis and display, In: "DNA Microarrays" edited by D. Bowtell and J. Sambrook, Cold Spring Harbor Laboratory (2002): 569-581.

[7] D. Spies, P.F. Renz, T.A. Beyer, C. Ciaudo, Comparative analysis of differential gene expression tools for RNA sequencing time course data, Briefings in Bioinformatics 20(1) (2019): 288–298.

[8] D. Palejev, Comparison of RNA-Seq Differential Expression Methods, Cybernetics and Information Technologies 17(5) (2017): 60-67.

[9] T. Wang, B. Li, C.E. Nelson, et al. Comparative analysis of differential gene expression analysis tools for single-cell RNA sequencing data. BMC Bioinformatics 20, 40, 2019.

[10] O. Georgieva, Iterative Clustering for Differential Gene Expression Analysis. In: Rojas, I., Valenzuela, O., Rojas, F., Herrera, L.J., Ortuño, F. (eds) Bioinformatics and Biomedical Engineering. IWBBIO 2022. Lecture Notes in Computer Science 13347. Springer, Cham., 2022.

[11] F.D. Smet, J. Mathys, K. Marchal, G. Thijs, B.D. Moor, and Y. Moreau, "Adaptive Quality-Based Clustering of Gene Expression Profiles," Bioinformatics 18 (2002): 735-746.

[12] A. P. Gasch, M.B. Eisen, Exploring the conditional coregulation of yeast gene expression through fuzzy k-means clustering, Genome Biology, 3(11), (2002): 1-22.

[13] O. Georgieva, F. Klawonn, E. Härtig, Fuzzy Clustering of Macroarray Data. In: Reusch, B. (eds) Computational Intelligence, Theory and Applications. Advances in Soft Computing 33. Springer, Berlin, Heidelberg, 2005.

[14] D. Kumar, U. Batra, Clustering algorithm for gene expression data. Int J Recent Res Asp. 4, (2017): 122-128.

[15] S. Srivastava, N. Joshi. Clustering techniques analysis for microarray data. Int J Comput Sci Mob Comput. 3 (2014); 359-364.

[16] M. Ester, H.-P Kriegel, J. Sander, X. Xiaowei, A density-based algorithm for discovering clusters in large spatial databases with noise. In: Proceedings of the Second International Conference on Knowledge Discovery in Databases and Data Mining, 226–231. AAAI Press, Portland, 1996.

[17] D. .Bottomly, N. A. R. Walter, J. E. Hunter et al. Evaluating Gene Expression in C57BL/6J and DBA/2J Mouse Striatum Using RNA-Seq and Microarrays. – PLoS ONE 6(3), e17820, 2011.