Improvement of K-Means Clustering Algorithm Performance in Gene Expression Data Analysis through Pre-Processing ....

José Arturo Molina Mora
University of Costa Rica
10 PUBLICATIONS **7** CITATIONS
SEE PROFILE

Fernando Mata
NutriScience
33 PUBLICATIONS **32** CITATIONS
SEE PROFILE

Diego A Bonilla Ocampo
Universidad Distrital Francisco José de Caldas
22 PUBLICATIONS **2** CITATIONS
SEE PROFILE

Life Science Informatics Publications

Research Journal of Life Sciences, Bioinformatics, Pharmaceutical and Chemical Sciences

Journal Home page http://www.rjlbpcs.com/

RJLBPCS

ISSN 2454-6348

Original Research Article

IMPROVEMENT OF K-MEANS CLUSTERING ALGORITHM PERFORMANCE IN GENE EXPRESSION DATA ANALYSIS THROUGH PRE-PROCESSING WITH PRINCIPAL COMPONENT ANALYSIS AND BOOSTING

Jose Arturo Molina Mora[1,*], Fernando Mata Ordoñez[2,3], and Diego Alexander Bonilla[4,5]

[1] Escuela de Matemática, Universidad de Costa Rica, San José, Costa Rica.

[2] Instituto Internacional de Ciencias del Ejercicio Físico y Salud, Murcia, España.

[3] NutriScience, Lisboa, Portugal.

[4] Línea de Investigación en Bioquímica y Biología Molecular, Universidad Distrital Francisco José de Caldas, Bogotá D.C., Colombia.

[5] Grupo de Investigación en Ciencias de la Actividad Física, el Deporte y la Salud, Universidad de Córdoba, Montería, Colombia.

ABSTRACT: High volume of data from recent high-throughput omics technologies and the study of particular diseases, such as cancer, represent a challenge to create new platforms able to analyze complex data sets, in order to extract biologically relevant information. The aim of this work is to analyze gene expression data from a public database of healthy and tumor tissues applying pre-processing with principal component analysis (PCA) and Boosting algorithm (BA), in order to increase the resolving power of a K-means clustering algorithm and to reduce noise in data during gene expression analysis. Data matrix was extracted from Princeton University Gene Expression Project, containing expression levels of 2000 genes taken from 62 different samples (healthy and tumor tissue). Clustering and pre-processing algorithms were executed in Weka 3.0. GenBank was used for gene annotation. After PCA, the resolving power of K-means model to cluster data, according to tissue and expression level, increased in 100% compared to a low capacity (54%) if not executed. Moreover, BA improved clustering in 87%, using only four genes; therefore, extraction of important biological information was possible. In conclusion, including data pre-processing using PCA augment the resolving power of a clustering algorithm. Furthermore, even with a discrete loss in precision, using BA allow to identify genes with the highest impact on cluster, making possible the extraction of crucial biological information.

KEYWORDS: Data Mining, K-Means, Principal Component Analysis, Boosting Algorithm, Cancer, Gene Expression.

***Corresponding Author: Prof. Jose Arturo Molina Mora**

Universidad de Costa Rica, 2060 San José, Costa Rica Telephone: (506) 2511-0000

* Email Address: jose.molinamora@ucr.ac.cr

1. INTRODUCTION

With the amount of data generated by the high-throughput technologies, bioinformatics has the challenge of bridging the gap between exponential data generation and low analytical capacity, a phenomenon called the "*Curse of Dimensionality*" [1,2]. In particular, several studies related to the cancer study have attempted to infer changes in gene expression in tumor tissue as compared to normal tissue, which involves the analysis of large and highly complex data sets to extract biologically relevant information. This complexity of data still requires analysis and computational-mathematical development, as an implicit tool for the extraction of non-trivial and prominent information [3]. In some studies, data mining techniques and algorithms have been used to filter irrelevant or redundant data, in order to use of other bioinformatics software or even mathematical-modeling afterwards. This leads to increased resolving capacity in analysis and predictions, as well as reliable associations according to context [1,4]. In detail, data mining is the process by which relevant and non-trivial patterns / information are extracted, analyzed and predicted from large data repositories [3]. The nature of this obtained data may be either quantitative or qualitative, besides of giving a great potential of utility that was not known. The process of data mining encompasses mathematical tools, statistics and algorithms; in fact, some authors consider mining as the properly use of tools to extract information within the process of discovery of knowledge in databases [5]. Data mining tasks include sorting, clustering, factor analysis, regression, and analysis of both association rules and sequential data. Sorting in data mining involves dividing the data set to assign defined categories or groups previously known, with the aim of finding models or functions that describe and distinguish classes for future predictions. Subsequently, clustering consists of grouping data into clusters (joined or disjoined), but different from sorting in terms categories of classification are not known [6]. The formation of the clusters is based on the similarity (metric or non-metric defined) between the elements of the set. In addition, since the group is obtained from the mining activity, researchers must give an interpretation of the formed groups. One of the most popular algorithms is the K-means, which aims to partition a set of *n* observations into *k* groups, in which each observation belongs to the group whose mean value is closest [1].

On the other hand, the factor analysis corresponds to a set of data reduction techniques (multivariate), which is used to explain the structure of correlations between the observed variables, in terms of a smaller number of unobserved variables called *factors*. Its main purpose is to find the underlying structure in a data matrix (*hidden factors*), to consider each factor as a dependent variable that is a function of the whole set of observed variables [7]. Thus, the central target is the summary and the reduction of data, taking into account that factors represent the linear combinations of the original variables with a minimum loss of information [6]. One of the most popular factor analysis methods is Principal Component Analysis (PCA). The algorithm corresponds to a method of information synthesis, with the reduction of variable size (the number of variables), which leads to the best plane (subspace) to visualize the cloud of points or data distribution. However, the new variables do not have the original dimensions, so the interpretation of the subsequent applied models may be uncertain. The algorithm pipeline starts from; i) an array of original data, ii) these are centered and reduced, iii) correlation matrix between the variables is established, and iv) the own vector and the characteristic value are achieved, to have the maximum dispersion. The new variables are constructed according to the order of importance in terms of the total variability that is collected from the original data [7]. In essence, we look for variables that are uncorrelated and linear combinations of the originals, collecting most of the information and variability of the data [6]. With the obtained values or weights, corresponding to each variable of the data matrix, the variables are combined to give rise to the first component (this component has an eigenvalue, which means how much of variance it explains). In a similar way, the other components are chosen, which are graphed as perpendicular planes to the previous components afterwards. The essential of this method is that each step provides with a reduction of the initial information, and therefore in few components the majority of information is collected. Eigenvalues are ordered and the components that provide the main information are set, so that we have an output that is the main component matrix, smaller in size than the original.

Currently, there are some assembly methods that combine a set of sorting tools to get the highest precision, in comparison to individual results. Within the methods for combining classifiers or assemblers, the most frequently used for resampling are Bagging and *Boosting* [8]. The *Boosting* method belongs to the well-known "*ensemble*" methods (combining), in which it is necessary to define how the set of classifiers was obtained (universe of hypotheses); for example, by a means of set of classifiers that works with several learning methods or what is called classifiers combination method. In general, the inclusion of *Boosting* methodology, in any learning set, is better than random prediction and classifiers such as decision trees and neural networks can be used. There are several *Boosting* modifications, including *adaBoosting, brownBoosting* and *logitBoosting.* In all cases, an output of binomial class represented as {-1, 1} is assumed. In the case of LogitBoost, this directly optimizes the logarithm of binomial likelihood (Bernoulli's likelihood criterion) and adjusts the logistic regression models with Newton's steps. The algorithm is based on the additive logistic regression model of the training data, which allows reducing the data through statistically significant and relevant selection. In this sense, the objective of this study is to perform a gene expression analysis of normal and tumor tissue with and without PCA to evaluate the resolving capacity of data clustering models, as well as to identify the particular expression of genes linked to cancer by applying *Boosting* algorithm.

2. MATERIAL AND METHODS

Data collection

The data consisted of expression levels of 2000 genes taken from 62 different samples, either from a tumor biopsy or from normal tissue, using microarrays. These public data are available at http://genomics-pubs.princeton.edu/oncology/.

Dimensionality Reduction

The data were grouped into two sets (without pre-defined categories), based on expression data and after evaluating their ability to discriminate between normal and tumor tissues using a K-means algorithm. In order to improve the resolution capacity, a pre-processing was performed using PCA and *Boosting*, to repeat the grouping using K-means (four best candidate genes were used in the case of *Boosting*). For these procedures the algorithms were executed in the Weka 3.0 software (http://www.cs.waikato.ac.nz/ml/weka/) (see Frank et al, 2016 for details). A pipeline was built to incorporate the clustering algorithms and Pre-processing (Figure 1).

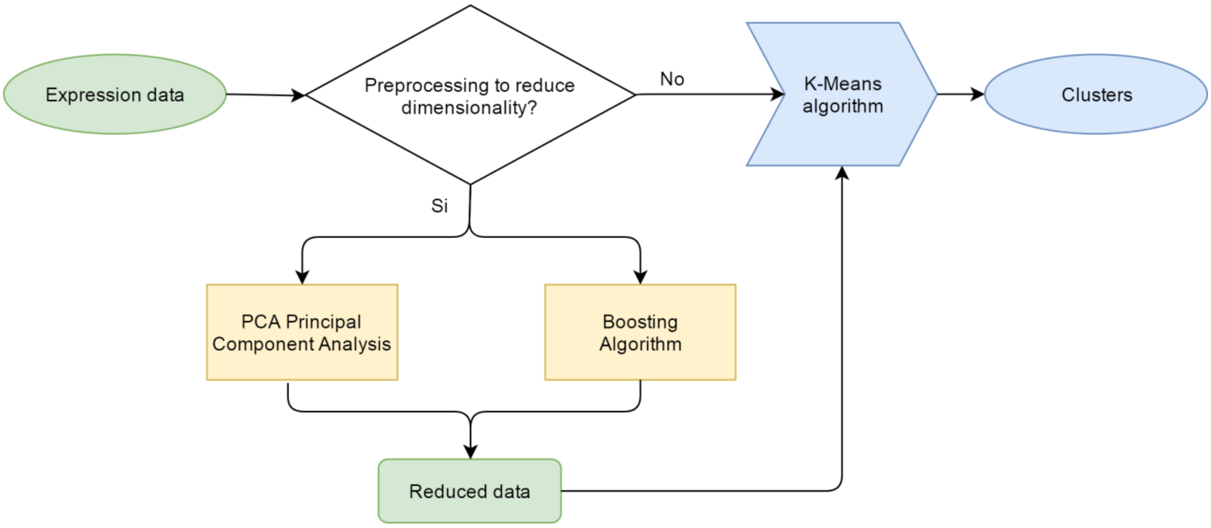

Figure 1. K-Means algorithm pipeline in Weka 3.0. Expression data was executed either with or without pre-processing, for reducing dimensionality, making possible the cluster generation in both cases.

Selection of differentially expressed genes in cancer

To identify differentially expressed genes in cancer versus normal tissue, the results from the Boosting algorithm were used to give biological interpretation to the selected genes and to evaluate their role in the development of this pathological condition. This interpretation was performed with the gene annotations available in the GenBank database (https://www.ncbi.nlm.nih.gov/genbank/).

3. RESULT AND DISCUSSION

Expression data grouping

After grouping data using the pipeline in Weka 3.0 (Figure 1), the performance of the K-means algorithm to group the expression data into tumor or normal condition was analyzed. After running the clustering, two groups were generated, one of 20 tissues (32%) and the second of 42 (68%); however, when re-sorting to each category (tumor or normal), the resolving capacity was relatively low, since the model does not discriminate adequately between normal tissues (green color in the first row) or tumor (red in the first row) (See second row in Figure 2).A desirable resolution of the algorithm should group normal and tumor tissues separately. In this data matrix, normal and malignant biopsies were ordered. But due to the low ability of the algorithm to adequately group tissues by types (only 54%); pre-processing of data was performed to increase discriminant power.

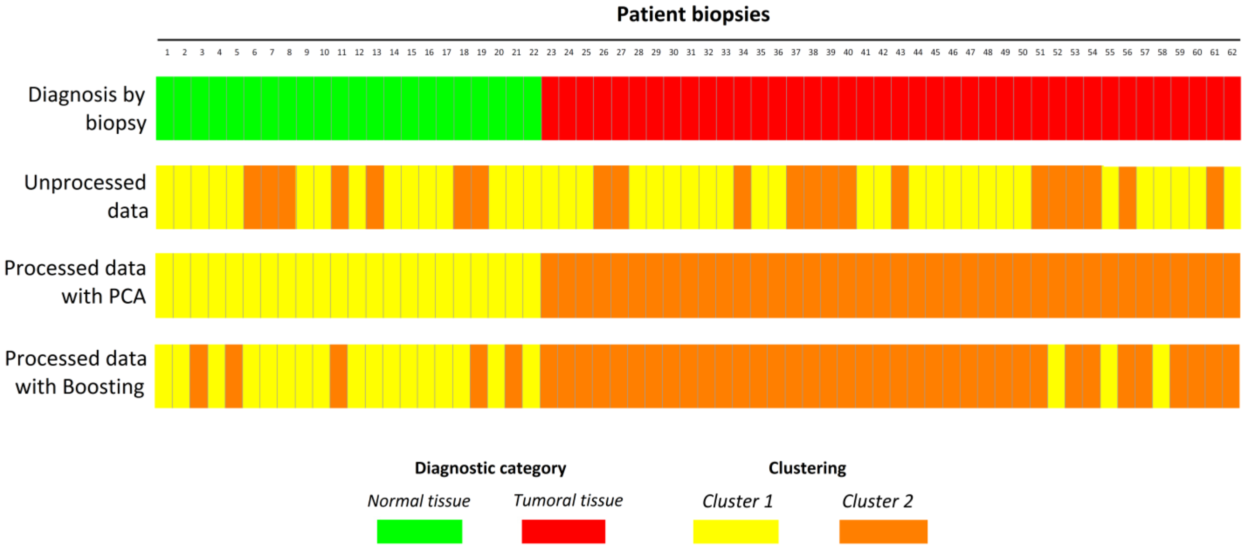

Figure 2. Analysis of expression data by K-means algorithm with or without pre-processing with PCA algorithm and *Boosting* algorithm.

Dimensionality reduction by PCA

After Principal component analysis (PCA) was performed, to eliminate irrelevant and redundant data, the number of attributes was reduced from 2000 genes to 25 components. Therefore, when the K-means algorithm workflow was applied again, the clustering model obtained two groups, one of 40 tissues (65%) and the other with 22 (35%). Comparing with the graph of diagnostic conditions for normal or tumor tissue, the maximum resolution capacity was obtained, since all elements of Cluster 1 (Figure 2, third row) correspond to normal tissues and elements of Cluster 2 (same row) are consistent with tumor tissues. Thus, the ability of the K-means algorithm, to cluster correctly, is improved after the application of the PCA (reaching 100%).

Reduction of dimensionality by Boosting Algorithm

In order to extract the genes with the greatest predictive power from the data matrix, a *logitBoosting* algorithm was applied, considering the binomial output {normal, tumoral}. To apply the algorithm, the conditions were specified to select the ten best genes that allow making the classification in the right categories (tumoral or normal). The results showed the following top ten predictor genes; Hsa.627, Hsa.1832, Hsa.1410, Hsa.6814, Hsa.3306, Hsa.1239, Hsa.11673, Hsa.8125 and Hsa.10047. These ten genes were used to create a data matrix, besides the class Tissue, to run again the K-means algorithm. After evaluating the clustering power with an even smaller number of genes, several gene deletions were performed to analyze the change in error of the K-means model. From those ten genes, four of them provided with most of the information to cluster, with a similar error in comparison to the full set of ten genes. The resulting genes were Hsa.1832, Hsa.6814, Hsa.3306 and Hsa.8125. Two groups were obtained after K-means execution using only these four genes, 20 tissues (32%) and 42 tissues (68%), with an error of 13% due to eight tissues with incorrect grouping (Figure 2, fourth row). In comparison to the model that uses full data without reducing, a better capacity of the algorithm to separate normal and malignant tissues was observed.

Table 1 shows the comparison of the accuracy obtained from the three models, where PCA pre-processing confers maximum precision. However, the use of only four genes from 2000 available (selected by *Boosting* algorithm) allowed an accuracy of 87%, as compared with the full data set without pre-processing in which a precision of 54% was achieved.

Table 1. COMPARISON OF ACCURACY WHEN APPLYING THE K-MEANS ALGORITHM IN THREE PRE-PROCESSING CONDITIONS

Model	**Accuracy (%)**	**Observations**
Full data with 2000 genes	54	Too much noise and redundancy Poor interpretation
PCA with 25 components	100	Total Noise Removal Poor interpretation
Boosting with 4 genes	87	Most noise elimination Possible to perform biological interpretation

High redundancy and noise of original data make necessary the use of pre-processing algorithms, especially because the low predictive power to get clusters and affected interpretation. In contrast, the increase in accuracy given by the PCA allows the total elimination of noise and extraction of relevant information, even though individual genes information is lost as components are generated, which may affect interpretation. On the other hand, when four genes from *Boosting* algorithm were selected for analysis, a smaller precision than PCA was achieved. However, this also included elimination of the majority of the noise, and an important reduction of biological entities that opened the possibility of making a suitable interpretation and biological explanation. This last is shown in the following section.

Characterization of differentially expressed genes

As mentioned before, the use of different pre-processing algorithms may improve accuracy of clustering techniques, but with some potential loss of biological interpretation. Therefore, a *Boosting* algorithm offers the chance to identify important genes and their possible link to experimental conditions after classification. Thus, it is essential to enrich information into the function of these genes and any direct association to normal and tumor tissues. To do so, GenBank database was used to determine the ontology of these genes and to make inferences of their possible role in both normal or tumor samples. In addition, a comparison of K-mean results was performed for the two types of tissues, by means of identifying the centroid value of the cluster, as an indirect measurement of the relative gene expression in both conditions. As a result, a relationship between the centroids from each K-means model, according to tissue, and the scientific literature was found (Table 2).

TABLE 2. COMPARISON OF GENES, ONTOLOGY AND CENTROIDS VALUES FROM THE K-MEANS MODEL.

Gene in data	**Hsa.1832**	**Hsa.6814**	**Hsa.3306**	**Hsa.8125**
ID in GenBank	J02854	1302	X12671	T71025
Function	Myosin Regulatory Light Chain 2	Collagen type XI alpha 2 chain	Heterogeneous nuclear ribonucleoprotein (hnRNP) core protein A1	Metallothionein 1G
Biological Importance	Apoptosis regulation Decrease level evades apoptosis	Positive regulation of the cell cycle Proliferation of cells	Move mRNA into the cytoplasm Proliferation	Architecture of the intestinal villi
Centroid of normal tissue	1047.8	62.87	350.45	2216.82
Centroid of tumoral tissue	188.6 (↓)	125.19 (↑)	817.70 (↑)	985.84 (↓)
Expression level according to bibliography	Decreased in colon cancer [9]	Increased marker in colon cancer (10)	Increased in cancer [6]	Decrease in colon carcinoma [9] Increased in normal tissue [11]

In detail, scientific literature of Hsa.1832 gene reports its decline in colon cancer, due to the role it plays in apoptosis. This reduction causes the cells to evade programmed cell death and proliferate, as was shown by Li & Wong in 2002 (9). Similarly, gene expression of metallothionein 1G (Hsa.8125) is diminished in tumor tissues, because of its link to intestinal villi architecture, particularly in the colon. Some authors consider metallothionein 1G as marker to monitor the progression of the disease [11].

Hsa.6814 gene corresponds to the Collagen type XI α-2 chain, which is a well-characterized marker of colon cancer and its expression is increased during the pathology progression, as it is related to the positive regulation of cell cycle and because of that favor proliferation of cells [10]. Finally, heterogeneous nuclear ribonucleoprotein (hnRNP) core protein A1, encoded by the Hsa.3306 gene, is increased in colon cancer, possibly due to its function of directing various messenger RNAs into the cytoplasm, which may lead to a direct influence in cell proliferation metabolism; notwithstanding, molecular mechanisms remains unclear [6]. These mentioned features of gene expression, and its relationship with cancer progression, support with some scientific evidence the algorithm's choice of the four genes, to cluster the evaluated tissues.

4. CONCLUSION

Analysis of gene expression data is a common task in bioinformatics, in which researchers must deal with a large amount of information and face certain difficulties. However, pre-processing analysis of raw data with well-designed algorithms is one of the most time-consuming and demanding process. In fact, our data showed no effective gene clustering without pre-processing, but resolution capacity of the algorithm was improved after PCA implementation. Nevertheless, since the components came from a linear combination of several inputs, the variable interpretability was lost. In compensation, even with a discrete loss in accuracy, the use of a *Boosting* algorithm allowed the identification of genes with the greatest impact in the cluster, in order to extract important biological information after ontology.

CONFLICT OF INTEREST

The authors declare that they have no competing financial or non-financial interest.

ACKNOWLEDGEMENTS

Authors gratefully acknowledge the Princeton University Gene Expression Project, without which the present study could not have been completed.

REFERENCES

[1] Bonilla E, Duval B, Hao J. Fuzzy Logic for Elimination of Redundant Information of Microarray Data. Genomics Proteomics Bioinformatics. 2008; 6(2): 61-73.

[2] Poggio T, Mhaskar H, Rosasco L, Miranda B, Liao Q. Why and When Can Deep--but Not Shallow--Networks Avoid the Curse of Dimensionality: a Review. arXiv preprint. 2016; arXiv:1611.00740.

[3] Bornholdt S. Less Is More in Modeling Large Genetic Networks. Science. 2015; 310: 449-451.

[4] Gough NR. New connections: Making discoveries in complex data sets. Sci Signal. 2016; 9(431): ec132.

[5] Fogel G. Computational intelligence approaches for pattern discovery in biological systems. Brief Bioinform. 2008; 9(4): 307-316.

[6] Li J, Wong L. Identifying good diagnostic gene groups from gene expression profiles using the concept of emerging patterns. Bioinformatics. 2002; 18(5): 725-734.

[7] Vineetha S, Shekara C, Idicula S. Gene regulatory network from microarray data of colon cancer patients using TSK-type recurrent neural fuzzy network. Gene. 2012; 506: 408–416.

[8] Reyes F, Hernández G, Calvo J. Métodos de clasificación de locutores utilizando clasificadores Boosting, 2011. Departamento de Ingeniería de sistema, Centro de Aplicaciones de Tecnologías de Avanzada (CENATAV), Ciudad de la Habana, Cuba. Reporte Técnico RT- 041 Serie Azul.

[9] Jin Y, Atkinson SJ, Marrs JA, Gallagher PJ. Myosin II Light Chain Phosphorylation Regulates Membrane Localization and Apoptotic Signaling of Tumor Necrosis Factor Receptor-1. J Biol Chem. 2001; 276: 30342-30349.

[10] Wang S, Chen H, Li F, Zhang D. Gene selection with rough sets for the molecular diagnosing of tumor based on support vector machines. In the Proceedings of the 2006 International Computer Symposium, Taiwan (China), 2006, pp. 1368-1373.

[11] Maglietta R, D'Addabbo A, Piepoli A, Perri F, Liuni S, Pesole G, et al. Selection of relevant genes in cancer diagnosis based on their prediction accuracy. Artif Intell Med. 2007; 40: 29-44.

[12] Frank E, Hall MA, Witten IH. The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques", Morgan Kaufmann, Fourth Edition, 2016. Available from: http://www.cs.waikato.ac.nz/ml/weka/