RESTRICTED CUR MATRIX DECOMPOSITION: A NOVEL TECHNIQUE FOR GENES SUBSET SELECTION IN PROBLEMS OF CLASSIFICATION OF CANCER TUMORS

Yunier Emilio Tejeda Rodríguez*[1]
* Universidad Central "Marta Abreu" de Las Villas, MES

ABSTRACT
In this paper we demonstrate the importance of restricted CUR matrix decomposition the selection of genes subset for the problem of classification of cancer tumors. We propose an algorithm that selects a genes subset of an unconventional way to the methods of selection features subset by filters. It tries to minimize the normalized Frobenius norm error by approximating the data matrix by a low rank matrix.
We demonstrate that normalized Frobenius norm error is a decreasing and bound succession, and converges to zero. We apply the proposed algorithm to a gene expression DNA microarray dataset in Colon cancer, setting the rank parameter k for the values 5, 7, 10, 13 and 20. Finally, we apply Principal Component Analysis to the subsets selected by the proposed algorithm for $k = 5$ in order to confirm the two classes of dataset: the sick class consisting of 40 patients and the healthy class consisting of 22 patients.

KEYWORDS: restricted CUR matrix decomposition, selection of gene subset, DNA microarray data set, Colon cancer, principal component analysis.

MSC: 62H25

RESUMEN
En este artículo demostramos la importancia de la descomposición matricial CUR restringida en la selección de subconjuntos de genes para el problema de clasificación de tumores cancerígenos. Con este fin, proponemos un algoritmo que selecciona un subconjunto de genes de una manera no convencional a los métodos por filtros. Este algoritmo intenta minimizar el error de la norma de Frobenius normalizado aproximando la matriz de datos mediante una matriz de bajo rango. Demostramos que el error de la norma de Frobenius normalizado es una sucesión decreciente y acotada, y converge a cero. Aplicamos el algoritmo propuesto a un conjunto de datos de microarreglos de ADN de expresión genética en cáncer de Colon. Para ello, se establecieron diferentes valores del parámetro de rango k: 5, 7, 10, 13 y 20. Finalmente, aplicamos el Análisis de Componentes Principales en los subconjuntos seleccionados por el algoritmo propuesto para k = 5. Los resultados del Análisis de Componentes Principales arrojaron la confirmación de las dos clases del conjunto de datos: la clase enferma que consta de 40 pacientes y la clase sana que consta de 22 pacientes.

PALABRAS CLAVES: descomposición matricial CUR restringida, selección de subconjunto de genes, conjunto de microarreglos de ADN, cáncer de Colon, análisis de componentes principales.

1. INTRODUCTION

The primary structure of the chromosome DNA chains containing all the genes of an organism, as well as all the components intervening in their regulation are revealed with the complete sequencing of genomes (Wheeler, D. A. ***et al.***, 2008.) This has allowed the development of technologies capable of analyzing all the identified elements of a genome in a single experiment (Miranda, J. and Bringas, R., 2008). One of these technologies is DNA microarrays: collections of DNA segments that are attached to a solid surface to be used in quantifying RNA or DNA levels in biological samples (Heller, M. J., 2002).
Applications of DNA microarray technology include gene discovery, disease diagnosis, drug discovery, and toxicology investigations. In disease diagnosis, this technology allows researchers to learn more about heart disease, mental illness, infectious diseases, and especially oncological diseases (Bednár, M., 2000).
In the case of the study of cancer, DNA microarrays allow us to distinguish between cancerous and non-cancerous samples, classify different types of cancer and identify subtypes of cancer that can progress aggressively (Hira, Z. and Gillies, D., 2015). In order to work with DNA microarrays, the genetic information obtained by this technology is processed and brought to an $n \times p$ matrix, where n is the number of patients and p is the number of genes studied. This matrix is called a DNA microarray dataset and, depending on the cancer study carried out, it is classified as a binary or multiclass DNA microarray dataset (Bolón-Canedo, V. ***et al.***, 2014).

[1] yunier@uclv.cu

A very common characteristic in DNA microarray data for cancer is its high dimensionality due to the number of genes studied, which ranges from thousands to tens of thousands of genes (Fan, J. and Li, R. 2006; Bolón-Canedo, V. ***et al.***, 2014). This high-dimensional classification problem is known as the "greater p smaller n" problem (Johnstone, I. and Titterington, D., 2009) and it includes other complexities such as the curse of dimensionality (Bellman, R. E., 1957) and the obtention of overfitted models (Everitt, B. S. and Skrondal, A., 2010); these complexities make classification an even bigger task. Imbalance and overlap among classes as well as the presence of missing, shifted and outlier data (Hambali, M. A., Oladele, T. O. and Adewole, K. S., 2020), are other problems that often appear in DNA microarray data for the Cancer. When faced with such challenges, the application of traditional methods in the search for a solution is difficult or impossible in some cases (Wang, N. N., 2009). Hence, it is necessary to look for other alternatives to deal with them.

Dimension reduction in DNA microarray data in cancer is an alternative to solve the high-dimensional classification problem presented by these data (Boulesteix, A., 2004). In order to solve this problem, the dimension reduction methods deal with transforming the data from a high-dimensional space to a low-dimensional space, so that this representation retains as much information as possible about them (Van Der Maaten, L., Postma, E. and Van den Herik, J., 2009). Such transformation is possible through features subset selection methods, which work by eliminating features that are irrelevant and redundant (Saeys, Y., Inza, I. and Larrañaga, P., 2007).

One of the simplest and fastest ways to select a subset of features are filter methods (Sánchez-Maroño, N., Alonso-Betanzos, A. and Tombilla-Sanromán, M., 2007), characterized by selecting the features of the data without involving any type of classification technique (Bolón-Canedo, V. ***et al***., 2014). The selection process of these methods consists of evaluating only the intrinsic properties of the data through statistical measures that include distance, information, dependency and consistency (Lavanya, C., Nandihini, M., Niranjana, R. and Gunavathi, C., 2014). Based on the intrinsic evaluation of the data, these methods are divided into univariate and multivariate.

Univariate filter methods are characterized by considering each feature separately (Hira, Z. and Gillies, D., 2015), among which are the formation of random subsets by random sampling (Parmigiani, G. ***et al***., 2003), the t-statistic test (Dai, J., Lieu, L. and Rocke, D., 2006) and Information Gain (IG) (Hall, M. and Smith, L., 1998). On the other hand, multivariate filter methods are capable of finding relationships between features (Hira, Z. and Gillies, D., 2015), the most widely used being Correlation Feature Selection (CFS) (Hall, M., 1999), the Fast Correlation-Based Filter (FCBF) (Yu, L. and Liu, H., 2003), ReliefF (Kononenko, I., 1994) and the minimum Relevance Maximum Redundancy (mRMR) (Peng, H., Long, F. and Ding, C., 2005).

In recent years, randomized algorithms have received much attention for the dimension reduction in large matrix problems. These algorithms refer to a class of random projection and random sampling algorithms recently developed to solve the least squares approximation and low rank matrix approximation problems, respectively (Mahoney, M. W., 2011; Kishore Kumar, N. and Schneider, J., 2016; Benjamin Erichson, N. ***et al.***, 2018). Among the random sampling algorithms are those that deal with the Column Subset Selection Problem, a highly topical area of research and theoretical and practical importance (Boutsidis, C., 2011). An example is the restricted CUR matrix decomposition or CX decomposition (Mahoney, M. W. and Drineas, P., 2009), a decomposition that has proven its usefulness as a dimension reduction technique.

In this paper we demonstrate the importance of restricted CUR matrix decomposition in the selection of genes subset for the problem of classification of cancer tumors. We propose an algorithm that selects a genes subset of an unconventional way to the methods of selection features subset by filters. It tries to minimize the normalized Frobenius norm error by approximating the data matrix by a low rank matrix. We demonstrate that normalized Frobenius norm error is a decreasing and bound succession, and converges to zero. We apply the proposed algorithm to a gene expression DNA microarray dataset in Colon cancer, setting the rank parameter k for the values 5, 7, 10, 13 and 20 corresponding to 70%, 75%, 80%, 85% and 90% of the variance explained by the principal components, respectively. For each of the five previous cases, the exact.num.random, top.scores, ortho.top.scores and highest.ranks methods derived from the restricted CUR matrix decomposition proposed by Mahoney, M. W. and Drineas, P. (2009) were applied. Finally, we apply the Principal Component Analysis to the subsets selected by the algorithm proposed for $k = 5$ in order to confirm the two classes of these dataset: the "sick class" consisting of 40 patients and the "healthy class" consisting of 22 patients.

The document is organized as follows: Section 2 describes the restricted CUR matrix decomposition and the ColumnSelect algorithm. Section 3 defines the normalized Frobenius norm error and details the RCURd proposed algorithm. Section 4 presents the DNA microarray dataset to be used. Sections 5 and 6 show the results and the discussion of the work, respectively. Finally, Section 7 presents the conclusions of the document.

2. METHODS

2.1. Restricted Cur Decomposition Matrix

Random sampling algorithms are a class of random algorithms that began to appear in the first decade of the 21st century to deal with very large matrix problems ranging from astronomy to genetics (Mahoney, M. W., 2011; Kishore Kumar, N. and Schneider, J., 2016; Benjamin Erichson, N. ***et al.***, 2018). An example of this is the Low Rank Matrix Approximation Problem, which consists of looking for a good approximation of an $m \times n$ data matrix A by a k low-rank matrix with $k \ll min(m,n)$ (Mahoney, M. W., 2011).

One of the low rank matrix approximation problems is the so-called Column Subset Selection Problem, which is defined as the selection of k columns from an $m \times n$ matrix A to form an $m \times k$ matrix C such that the residual $\|A - P_C A\|_F$ is minimal for all possible choices $\binom{n}{k}$ of matrix C; where $P_C = \mathrm{CC}^+$ denotes the projection onto the generated k-dimensional space by the columns of C and $\|\cdot\|_F$ denotes the Frobenius norm (Drineas, P., Mahoney, M. W. and Muthukrishnan, S., 2006).

Among the low-rank matrix approximation random sampling algorithms that deal with the Column Subset Selection Problem is the restricted CUR matrix decomposition or CX decomposition (Mahoney, M. W. and Drineas, P., 2009), that allows the approximation of an input matrix A through the product of two matrices C and X , where C contains some columns of matrix A and X is a matrix that guarantees aforementioned approximation (Drineas, P., Mahoney, M. W. and Muthukrishnan, S., 2006a; Drineas, P., Mahoney, M. W. and Muthukrishnan, S., 2006b; Boutsidis, C., Mahoney, M. W. and Drineas, P., 2008; Drineas, P., Mahoney, M. W. and Muthukrishnan, S. , 2008; Boutsidis, C., Mahoney, M. W. and Drineas, P., 2009; Mahoney, M. W. and Drineas, P., 2009; Boutsidis, C., 2011; Papailiopoulos, D., Kyrillidis, A. and Boutsidis, C ., 2014.)

2.2. ColumnSelect ALGORITHM

In 2009, Mahoney and Drineas proposed a restricted CUR matrix decomposition, whose criterion for choosing the columns that form the matrix C consists of an importance factor defined by $\pi_j = \frac{1}{k}\sum_{i=1}^{k}\left(v_j^i\right)^2, \forall j = 1,\ldots,n$ where v_j^i is the j-th component of the i-th right singular vector of A. To build matrix C the authors created the ColumnSelect algorithm that takes as input any $m \times n$ matrix A, a rank parameter k and an error parameter ϵ.

The selection process of the c columns that form the matrix C in the ColumnSelect algorithm (table 1) begins with the computation of the largest k right singular vectors of A (line 1). Then, the j-th normalized importance factor is computed, π_j, for each of the columns (line 2). Subsequently, the j-th selection probability of each column is computed as p_j, the minimum between 1 and $c\pi_j$. Select the j-th column when its selection probability p_j is greater than or equal to the probability obtained by a uniform distribution with parameter n (line 3). The algorithm stops when it reaches the number of c selected columns, even if there are more p_j greater than or equal to the uniform probability (line 4).

Table 1: Description of the ColumnSelect algorithm.

Input: Data matrix $A_{m\times n}$, rank parameter k and error parameter ϵ.
Output: Matrix with a few columns of $A_{m\times n}$: $C_{m\times c}$ with $c \ll n$.
1: Compute $v^1, v^2, \ldots, v^k$.
2: Compute the normalized importance factors according to π_j.
3: Keep the jth column of A with probability $p_j = min(1,\ c\pi_j)$, for all $j = 1,\ldots,n$ where $c = O\left(k\frac{\log k}{\varepsilon^2}\right)$.
4: Return the matrix C consisting of the selected columns of A.

An implementation of the ColumnSelect algorithm is found in the rCUR package (Bodor, A., Csabai, I, Mahoney, M. W. and Solymosi, N., 2012) of the R software implemented by Bodor, A. and Solymosi, N. (2011). The methods random, exact.num.random, top.scores, ortho.top.scores, and highest.ranks appear in this package and provide the same precision as the ColumSelect algorithm.

The most important theoretical result supporting the ColumnSelect algorithm states that this selection of columns satisfies the inequality $\|A - P_C A\|_F \leqslant \left(1 + \frac{\epsilon}{2}\right)\|A - A_k\|_F$ with a probability of 99% at least, where $P_C A$ denotes the projection matrix on the column space generated by C and A_k is the matrix of rank

k closest to A in Frobenius norm (Drinea, P., Mahoney, M. W. and Muthukrishnan, S., 2008). This way the result guarantees that if A is a matrix close to another of rank k, then the subspace generated by the columns of A is close to the subspace generated by the columns of C, with high probability.

3. PROPOSED ALGORITHM

3.1. Normalized Frobenius Norm Error

The normalized Frobenius norm error $\frac{\|A-P_C A\|_F}{\|A-A_k\|_F}$ can be considered as a function of the number of selected columns c, this is: $\theta(c) = \frac{\|A-P_{C(c)}A\|_F}{\|A-A_k\|_F}$, whose domain is the set of natural numbers $1, 2, \dots, p$ and whose image is the set of positive real numbers. Therefore, this error is a succession $\{\theta(c)\}_{c=1}^{p}$ and it can be shown that it converges to zero. In order to prove this we will use a result from the theory of numerical successions: Every decreasing and bounded succession converges towards its lower end.
In order to prove that the succession is decreasing, let us consider, without loss of generality, the matrices $C(c)$ and $C(c+1)$ formed by the c and $(c+1)$ columns with higher importance factors of matrix A, respectively. Let us denote the matrices: (i) $C(c)$ by $\left(a_{ij}\right)_{m\times c}$ and $C(c+1)$ by $\left(a_{ij}\right)_{m\times(c+1)}$; (ii) $C^+(c)$ by $\left(b_{ij}\right)_{c\times m}$ and $C^+(c+1)$ by $\left(b_{ij}\right)_{(c+1)\times m}$; (iii) $C(c)\cdot C^+(c)$ by $\left(\sum_{k=1}^{c} a_{ik}\cdot b_{kj}\right)_{m\times m}$ and $C(c+1)\cdot C^+(c+1)$ by $\left(\sum_{k=1}^{c+1} a_{ik}\cdot b_{kj}\right)_{m\times m}$.
Let's show that $\theta(c) > \theta(c+1)\ \forall c \geq 1$.

$0 = \|A\|_F^2 - \|A\|_F^2$ (Nonnegative Frobenius norm)

$= \left\|A - P_{C(c)}A + P_{C(c)}A\right\|_F^2 - \left\|A - P_{C(c+1)}A + P_{C(c+1)}A\right\|_F^2$ (Addition and substraction of $P_{C(c)}A$ and $P_{C(c+1)}A$)

$\leq \left\|A - P_{C(c)}A\right\|_F^2 + \left\|P_{C(c)}A\right\|_F^2 - \left(\left\|A - P_{C(c+1)}A\right\|_F^2 + \left\|P_{C(c+1)}A\right\|_F^2\right)$ (Triangle inequality)

$= \left\|A - P_{C(c)}A\right\|_F^2 + \|C(c)C^+(c)A\|_F^2 - \left\|A - P_{C(c+1)}A\right\|_F^2 - \|C(c+1)C^+(c+1)A\|_F^2$ (Definition of $P_{C(c)}A$)

$\leq \left\|A - P_{C(c)}A\right\|_F^2 + \|C(c)C^+(c)\|_F^2\cdot\|A\|_F^2 - \left\|A - P_{C(c+1)}A\right\|_F^2 - \|C(c+1)C^+(c+1)\|_F^2\cdot\|A\|_F^2$ (1)

$= \left\|A - P_{C(c)}A\right\|_F^2 + \|C(c)C^+(c)\|_F^2\cdot\|A\|_F^2 - \left\|A - P_{C(c+1)}A\right\|_F^2 - [\|C(c)C^+(c)\|_F^2 + \mathcal{K}]\cdot\|A\|_F^2$ (2)

$= \left\|A - P_{C(c)}A\right\|_F^2 + \|C(c)C^+(c)\|_F^2\cdot\|A\|_F^2 - \left\|A - P_{C(c+1)}A\right\|_F^2 - \|C(c)C^+(c)\|_F^2\cdot\|A\|_F^2 - \mathcal{K}\cdot\|A\|_F^2$ (3)

$= \left\|A - P_{C(c)}A\right\|_F^2 - \left\|A - P_{C(c+1)}A\right\|_F^2 - \mathcal{K}\cdot\|A\|_F^2$ (Reducing similar terms)

Inequality (1) is obtained by the Cauchy-Schwarz inequality while equalities (2) and (3) are obtained by expanding $P_{C(c+1)}A$ as a function of $P_{C(c)}A$ and the distributive property, respectively.
Then, $0 < \left\|A - P_{C(c)}A\right\|_F^2 - \left\|A - P_{C(c+1)}A\right\|_F^2 - \mathcal{K}\cdot\|A\|_F^2$ and multiplying both members of the inequality by $\frac{1}{\|A-A_k\|_F^2}$ we get $0 < \theta^2(c) - \theta^2(c+1) - \frac{\mathcal{K}\|A\|_F^2}{\|A-A_k\|_F^2}$.
Therefore, $0 < \frac{\mathcal{K}\cdot\|A\|_F^2}{\|A-A_k\|_F^2} < \theta^2(c) - \theta^2(c+1)$ where $\mathcal{K} = \sum_{i=1}^{m}\sum_{j=1}^{m}\left(a_{i(c+1)}\cdot b_{(c+1)j}\right)^2 + \sum_{i=1}^{m}\sum_{j=1}^{m} 2a_{i(c+1)}\cdot b_{(c+1)j}\cdot\left(\sum_{k=1}^{c} a_{ik}\cdot b_{kj}\right)$.
Finally, $\theta^2(c) > \theta^2(c+1)\ \forall c \geq 1$ therefore, $\theta(c) > \theta(c+1)\ \forall c \geq 1$.
In order to show that the succession is bounded, we start with the inequality $\|A - P_C A\|_F \leqslant \left(1 + \frac{\epsilon}{2}\right)\|A - A_k\|_F$. Solving $\|A - A_k\|_F$ on the left side of the inequality, we get an upper bound for the normalized Frobenius norm error $\frac{\|A-P_C A\|_F}{\|A-A_k\|_F}$ depending on the error parameter $1 + \frac{\epsilon}{2}$. On the other hand, $\frac{\|A-P_C A\|_F}{\|A-A_k\|_F} \geq 0$ since the Frobenius norm for a matrix A is defined by $\|A\|_F = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{p}\left(a_{ij}\right)^2}$.
Therefore, the succession $\{\theta(c)\}_{c=1}^{p}$ is upper bounded by zero and lower bounded by $1 + \frac{\epsilon}{2}$.
Finally, the succession $\{\theta(c)\}_{c=1}^{p}$ is decreasing and bounded, and converges towards its lower end, that is, $\lim_{c\to p}\theta(c) = 0$.

3.2. RCURd Algorithm

Restricted **CUR d**ecomposition (**RCURd**) seeks to select a subset of genes in a manner not conventional to filter feature subset selection methods by minimizing the normalized Frobenius norm error by approximating the data matrix A by a low rank matrix C. This gene subset selection differs from the selection process by filter methods, which select gene subsets by observing only the intrinsic features of

the data by statistical measures. Despite the differences between both methods, the two share one thing in common, neither involves the use of sorting techniques in the gene subset selection process.
The gene subset selection process in the RCURd algorithm (Figure 1), begins with the computation of the best approximation matrix of rank k by the Singular Value Decomposition (Golub, G. H. and Van Loan, C. F., 1996) (step 1). Then, the c-th terms are computed: $\theta(c) = \frac{\|A - P_{C(c)}A\|_F}{\|A - A_k\|_F}$ of succession $\{\theta(c)\}_{c=1}^{p}$ (step 2). Subsequently, the behavior of the succession is analyzed $\{\theta(c)\}_{c=1}^{p}$ when c grows to p looking for stability at zero value (step 3). The subset of genes selected by RCURd is $C = \min_{1 \le c \le p}\{C(c): \theta(c) = 0\}$ (step 4). Table 2 shows the description of the RCURd algorithm.

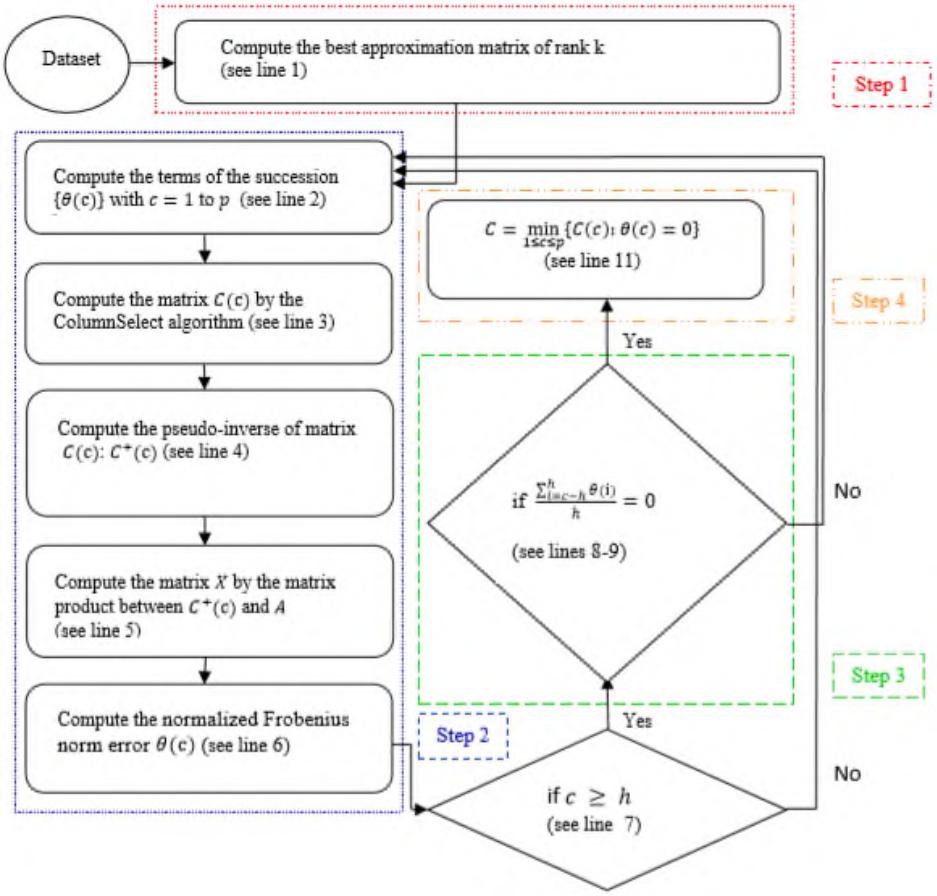

Figure 1: Description of the RCURd algorithm steps.

Table 2: Description of the RCURd algorithm.

Input: DNA microarray dataset $A_{n \times p}$ with $n \ll p$.
Output: Subset of genes selected $C_{n \times c}$ with $c \ll p$.
1: Compute the matrix A_k by the Singular Value Decomposition
2: **for** c = 1 **to** p **do**
3: Compute the matrix C(c) by the ColumnSelect algorithm
4: Compute the pseudo-inverse of matrix C(c), denoted by C^+(c)
5: Compute the matrix X by the matrix product between C^+(c) and A
6: Compute the normalized Frobenius norm error $\theta(c) = \frac{\|A - C(c) \cdot X\|_F}{\|A - A_k\|_F}$
7: **if** c ≥ h **then** {
8: **if** $\frac{\sum_{i=c-h}^{h} \theta(i)}{h} = 0$ **then** {
9: **break**
10: }
11: Select the gene subset of A by the matrix $C = C(c - h)$
12: }
13: **end**

4. DATASET

We worked with the gene expression DNA microarray dataset in Colon cancer (Alon, U. ***et al.***, 1999) available in the public international repository (Bio-Medical Dataset, 2017). This dataset contains the expression levels of 2,000 genes for 62 patients divided into two classes; the sick class, made up of 40 patients with colon cancer and the healthy class, made up of 22 patients. When analyzing the dimension of this set and the sample sizes of the two classes, it can be concluded that the data set presents the problem "greater p smaller n" and the imbalance between the classes with an imbalance rate of 1.82.

5. RESULTS

The research results were obtained using the R-3.4.3 software (R Core Team, 2017) on a computer running Windows 10 Pro 64-bit, 8 GB RAM, an Inter(R) processor Core(TM) i3-6100 and 1000GB capacity. In addition, we worked with the RStudio integrated development environment to implement the DCURr algorithm using the Matrix (Bates, D. and Maechler, M., 2017), MASS (Venables, W. N. and Ripley, B. D., 2002) and rCUR packages.
In order to study the convergence of the RCURd algorithm, the range parameter k was set for the values 5, 7, 10, 13 and 20 corresponding to 70%, 75%, 80%, 85% and 90% of the variance explained by the principal components, respectively. For each of the previous cases, the exact.num.random, top.scores, highest.ranks and ortho.top.scores methods were used. The results showed that the number of genes c in the subset selected by the RCURd algorithm turned out to be the number of samples in the colon cancer set, n = 62, since from this number the normalized Frobenius norm error starts to stabilize equal to zero. Figure 2 to Figure 6 show these results.

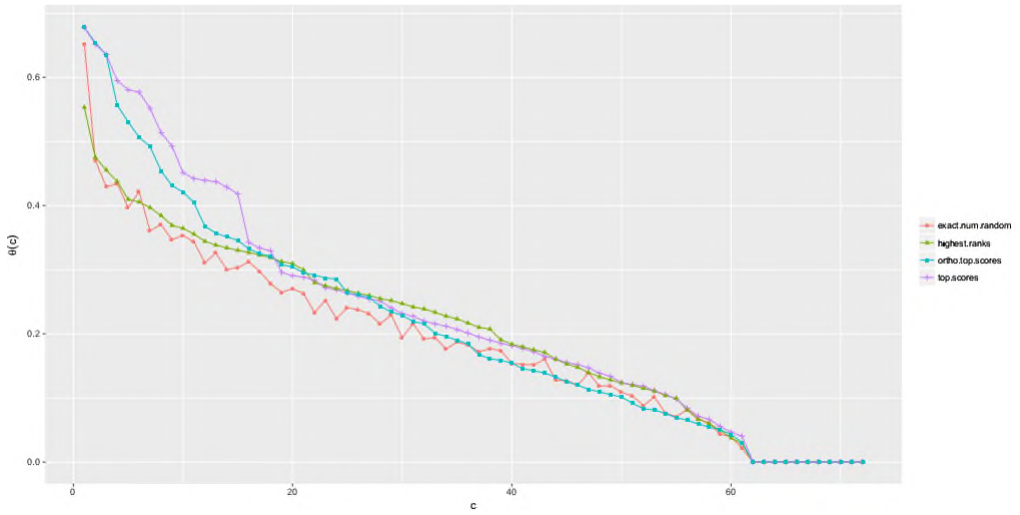

Figure 2: Convergence results of the RCURd algorithm for the range parameter k = 5.

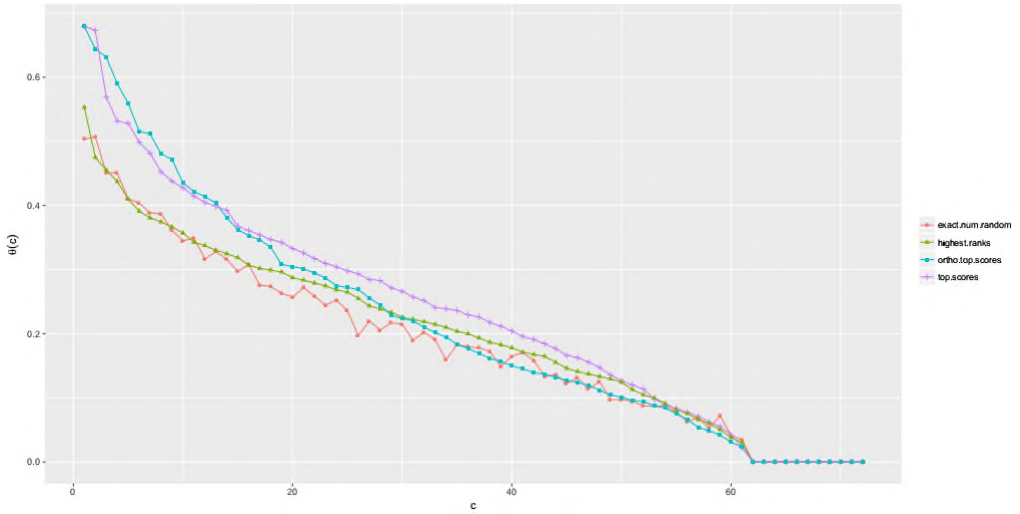

Figure 3: Convergence results of the RCURd algorithm for the range parameter k = 7.

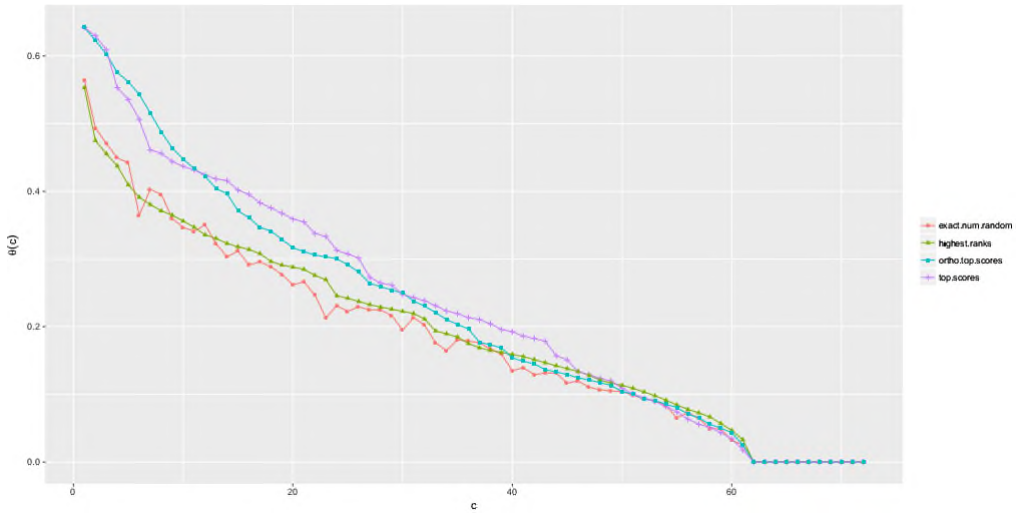

Figure 4: Convergence results of the RCURd algorithm for the range parameter k = 10.

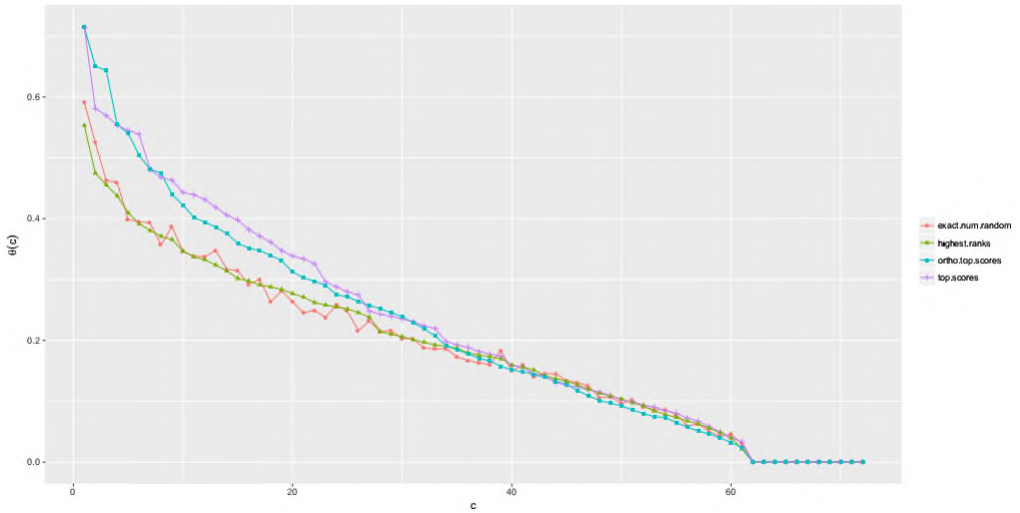

Figure 5: Convergence results of the RCURd algorithm for the range parameter k = 13.

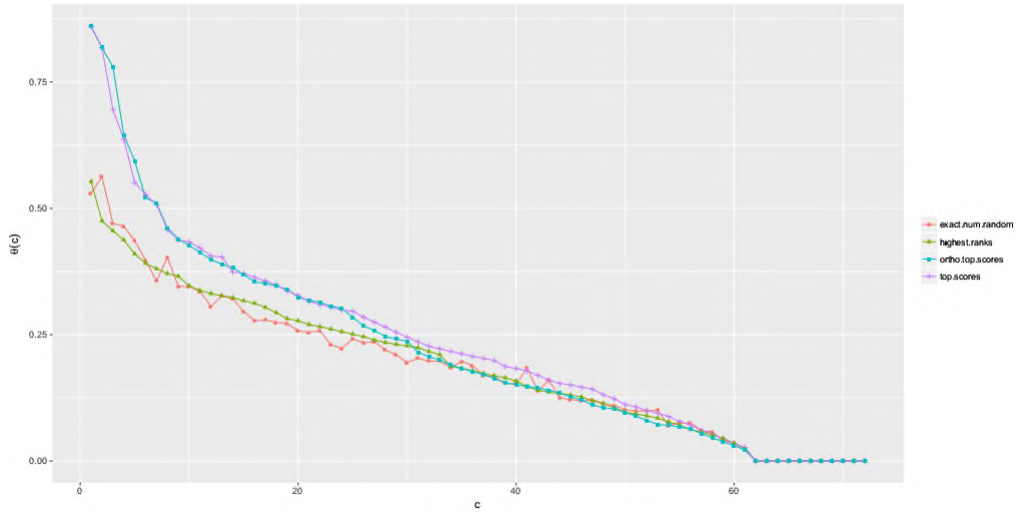

Figure 6: Convergence results of the RCURd algorithm for the range parameter k = 20.

6. DISCUSSION

Taking into account the results of the convergence of the RCURd algorithm, it was decided to determine the presence of genes that the selected subsets have in common. For this, a Venn diagram was made in each of the studies (k = 5, k = 7, k = 10, k = 13 and k = 20) whose results are shown in Figure 7 to Figure 11. After analyzing these figures, it was concluded that similar results were obtained in the five studies, resulting in the subsets selected by RCURd with top.scores and RCURd with highest.ranks having the highest presence of genes in common. To make the Venn diagram, the VennDiagram package (Hanbo, C., 2018) of the R software was used.

Figure 7: Venn diagram for the range parameter k = 5.

Figure 8: Venn diagram for the range parameter k = 7.

Figure 9: Venn diagram for the range parameter k = 10.

Figure 10: Venn diagram for the range parameter k = 13.

Figure 11: Venn diagram for the range parameter k = 20.

It is known that the dataset under study contains the expression levels of 2000 genes for 62 patients divided into two classes (sick and healthy); Principal Component Analysis was applied to the subsets of genes selected by RCURd algorithm seeking to confirm these two classes. For this, the stats package (R Core Team, 2017) of the R software was used to obtain the results. Given that the Venn diagrams were

similar for the five studies, it was decided to work with the subsets selected by RCURd for the range parameter k = 5.
Figures 12 to 15 show the scatterplot scores and the scatterplot loadings of the principal components for the gene subsets selected by the RCURd algorithm with the exact.num.random, highest.ranks, ortho.top.scores and top.scores methods, respectively. The results of the principal components showed that in Figure 12 it was not possible to confirm the sick and healthy classes in the subset of genes selected by RCURd with exact.num.random, despite the fact that the first two components explained 57.63% of the variance. This is mainly due to the presence of the imbalance between the classes manifested by this set. The same thing happened in figure 13 with the subset of genes selected by RCURd using highest.ranks whose first two components explained 67.25% of the variance. Notwithstanding this, it was possible to agglomerate the 62 selected genes into three groups of genes. On the other hand, in figure 14, an attempt to separate the classes in the subset of genes selected by RCURd with ortho.top.scores was observed despite the fact that the first two components explained less than 30% of the variance. Finally, in figure 15, it was observed how the subset of genes selected by RCURd with top.scores was able to confirm the two classes and group the 62 selected genes into two groups of genes (see Appendix I). Classifying a total of 40 patients with colon cancer 36 patients for 90% and a total of 22 healthy patients 10 patients for 45.45%, respectively.

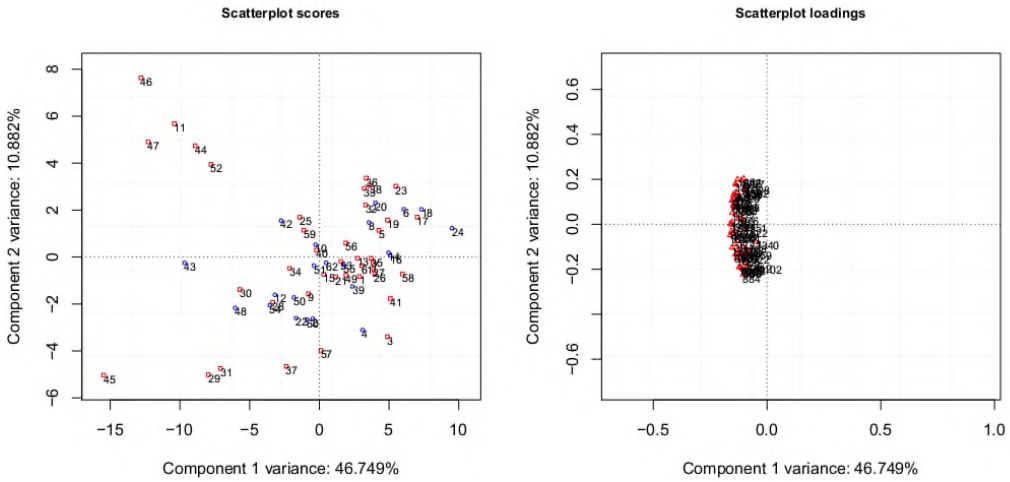

Figure 12: Results of the principal components in the subset of genes selected by RCURd with exact.num.random for the range parameter k = 5.

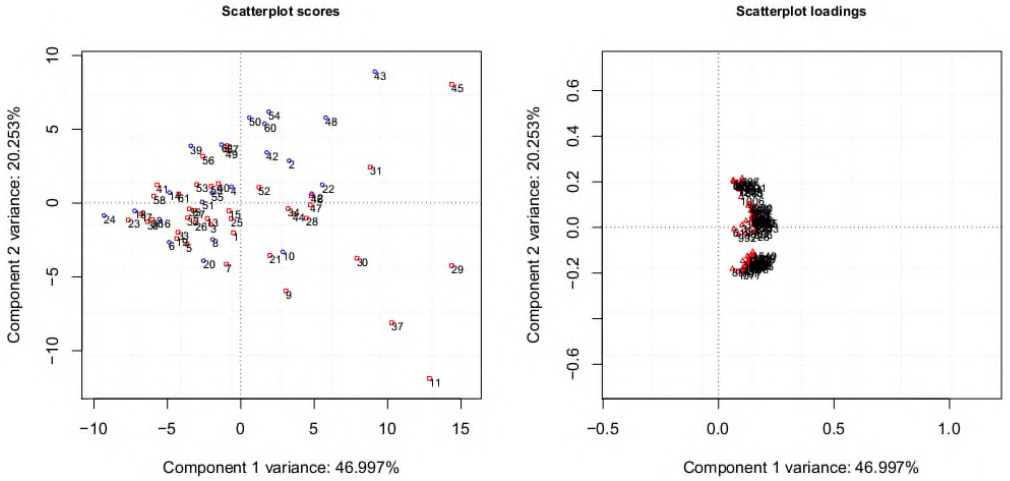

Figure 13: Results of the principal components in the subset of genes selected by RCURd with highest.ranks for the range parameter k = 5.

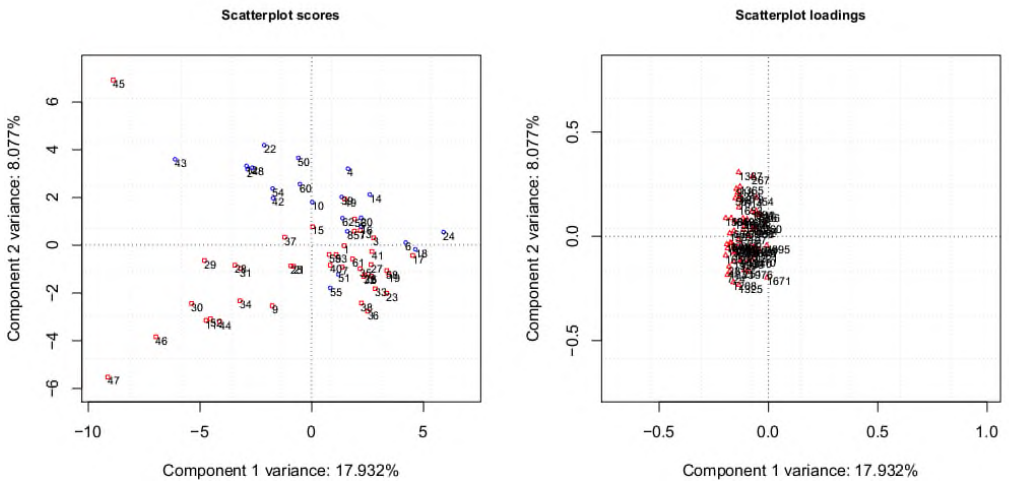

Figure 14: Results of the principal components in the subset of genes selected by RCURd with ortho.top.scores for the range parameter k = 5.

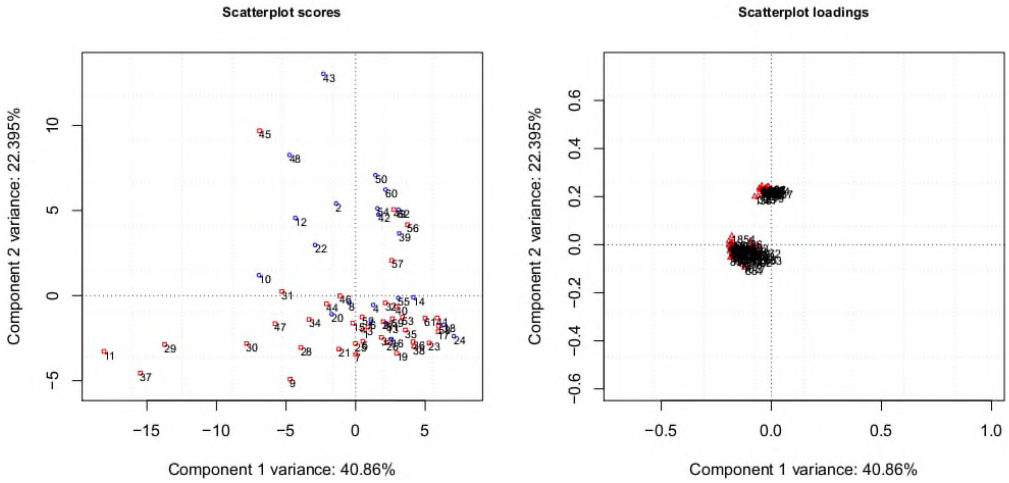

Figure 15: Results of the principal components in the subset of genes selected by RCURd with top.scores for the range parameter k = 5.

7. CONCLUSIONS

In this work, a novel algorithm, called RCURd, was proposed for gene subset selection in cancer tumor classification problems. This algorithm selects a subset of genes in a non-conventional way to features subset selection filter methods. It tries to minimize the normalized Frobenius norm error by approximating the data matrix by a low rank matrix. The error was considered as a succession depending on the number of genes and it was shown that it is decreasing and bounded, and converges to zero.
In order to study the convergence of the RCURd algorithm, the Colon cancer DNA microarray dataset containing the expression levels of 2000 genes for 62 patients divided into two classes (sick and healthy) was used. To do this, the range parameter k was set equal to 5, 7, 10, 13 and 20; using on each of these values the methods exact.num.random, top.scores, ortho.top.scores and highest.ranks derived from the ColumnSelect algorithm. The results of the convergence of the RCURd algorithm showed that the number of samples in the selected subset turned out to be 62, since from this number the Frobenius norm error normalized equal to zero began to stabilize. Finally, Principal Component Analysis was applied to the subsets selected by the RCURd algorithm with the exact.num.random, top.scores, ortho.top.scores, and highest.ranks methods for the rank parameter k = 5. The results of the principal components showed that in the subset of genes selected by the RCURd algorithm with the top.scores method, it was possible to confirm the two classes with the first two components and to group the selected genes into two groups.

ACKNOWLEDGMENTS: The author wishes to thank the contributions of Dr. Valia Guerra Ones, Dr. Jesús Eladio Sánchez García and Lic. Lorenzo Antonio Pérez Carballo.

RECEIVED: AUGUST, 2022.
REVISED: JANUARY , 2023.

REFERENCES

[1] ALON, U., BARKAI, N., NOTTERMAN, D. A., GISH, K., YBARRA, S., MACK, D., and LEVINE, A. J. (1999): Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. **Proceedings of the National Academy of Sciences**, 96, 6745-6750.
[2] BATES, D. and MAECHLER, M. (2017): Matrix: Sparse and Dense Matrix Classes and Methods. **R package version 1.2-12**. https://CRAN.R-project.org/package=Matrix.
[3] BEDNÁR M. (2000): DNA microarray techonology and application. **Med Sci Monit**. Jul-Aug; 796-800. PMID: 11208413.
[4] BELLMAN R. E. (1957): **Dynamic Programming**. Princeton University Press, Princeton.
[5] BENJAMIN ERICHSON, N., VORONIN, SERGEY, BRUNTON, STEVEN L. and NATHAN KUTZ, J. (2018): Randomized Matrix Decompositions using R. **Journal of Statistical Software**, VV, http://www.jstatsoft.org.
[6] Bio-Medical Dataset. http://datam.i2r.a-star.edu.sg/datasets/krbd **Consulted** September, 2017.
[7] BOLÓN-CANEDO, V., SÁNCHEZ-MAROÑO, N., ALONSO-BETANZOS, A., BENÍTEZ, J. and HERRERA, F. (2014): A review of microarray datasets and applied feature selection methods. **Information Sciences**, 282, 111–135.
[8] BODOR, A., CSABAI, I, MAHONEY, M. W. and SOLYMOSI, N. (2012): rCUR: an R package for CUR matrix decomposition. **BMC Bioinformatics**, 13, 1-6.
[9] BODOR, A. and SOLYMOSI, N. (2011): rCUR: CUR decomposition package. **R package version 1.0.** http://CRAN.R-project.org/package=rCUR.
[10] BOULESTEIX, A. (2004): PLS Dimension reduction for classification with microarray data. **Statistical Applications in Genetics and Molecular Biology**, 3, 1-33.
[11] BOUTSIDIS, C. (2011): **Topics in Matrix Sampling Algorithms**. PhD thesis, arXiv:1105.0709v1 [cs.DS] 4 May 2011.
[12] BOUTSIDIS, C., MAHONEY, M. W. and DRINEAS, P. (2008): On selecting exactly k columns from a matrix. **Submitted for publication**.
[13] BOUTSIDIS, C., MAHONEY, M. W. and DRINEAS, P. (2009): An Improved Approximation Algorithm for the Column Subset Selection Problem. In Proceedings of the Twentieth Annual ACM SIAM Symposium on Discrete Algorithms, 968-977. Society for Industrial and Applied Mathematics.
[14] DAI, J., LIEU, L. and ROCKE, D. (2006): Dimension Reduction for Classification with Gene Expression Microarray Data. **Statistical Applications in Genetics and Molecular Biology**, 5, https://doi.org/10.2202/1544-6115.1147.
[15] DRINEAS, P., MAHONEY, M. W. and MUTHUKRISHNAN, S. (2006): Polynomial time algorithm for column-row based relative-error low-rank matrix approximation. **DIMACS TR: 2006-04**, 1-15.
[16] DRINEAS, P., MAHONEY, M. W. and MUTHUKRISHNAN, S. (2006): Subspace Sampling and Relative-Error Matrix Approximation: Column-Based Methods. In **Approximation, Randomization and Combinatorial Optimization. Algorithms and Techniques**, 316-326. Springer, Berlin, Heidelberg.
[17] DRINEAS, P., MAHONEY, M. W. and MUTHUKRISHNAN, S. (2008): Relative-error CUR matrix decompositions. **SIAM Journal on Matrix Analysis and Applications**, 30, 844-881.
[18] EVERITT B.S. and SKRONDAL A. (2010): **Cambridge Dictionary of Statistics**, Cambridge University Press, Cambridge.
[19] FAN, J., and LI, R. (2006). Statistical challenges with high dimensionality: Feature selection in knowledge discovery. **arXiv preprint** math/0602133.
[20] HALL, M. and SMITH, L. (1998). Practical feature subset selection for machine learning, **Comput. Sci**. 98, 181–191
[21] GOLUB, G. H. and VAN LOAN C. F. (1996): **Matrix Computations**. Johns Hopkins University Press, Baltimore.
[22] HALL., M. (1999): **Correlation-Based Feature Selection for Machine Learning**. PhD thesis, Citeseer.
[23] HAMBALI, M.A., OLADELE, T.O., and ADEWOLE, K. S. (2020): Microarray Cancer Feature Selection: Review, Challenges and Research Directions. **International Journal of Cognitive Computing in Engineering**, 1, 78-97.
[24] HANBO C. (2018): VennDiagram: Generate High-Resolution Venn and Euler Plots. **R package version 1.6.20.** https://CRAN.R-project.org/package=VennDiagram.

[25] HELLER, M. J. (2002): DNA microarray techonology: devices, systems, and applications. **Annu Rev Biomed Eng**; 4:129-53. doi: 10.1146/annurev.bioeng.4.020702.153438. Epub 2002 Mar 22. PMID: 12117754.
[26] HIRA, ZENA and GILLIES, DUNCAN. (2015): A Review of Feature Selection and Feature Extraction Methods Applied on Microarray Data. **Advances in Bioinformatics**, 1-13. 10.1155/2015/198363.
[27] JOHNSTONE, IAIN and TITTERINGTON, D. (2009): Statistical challenges of high-dimensional data. **Phil. Trans. R. Soc. A**, 367, 4237-4253. 10.1098/rsta.2009.0159.
[28] KISHORE KUMAR N. and SCHNEIDER, J. (2016). Literature survey on low rank approximation of matrices, **National Board of Higher Mathematics**, India. 1-30.
[29] KONONENKO, I. (1994): Estimating attributes: analysis and extensions of relief, in: **Machine Learning: ECML-94**, Springer, 171–182.
[30] LAVANYA, C., NANDIHINI, M., NIRANJANA, R. and GUNAVATHI, C. (2014): Classification of microarray data based on feature selection method. **International Journal of Innovative Research in Science, Engineering and Technology**, 3, 1261-1264.
[31] MAHONEY, M.W. (2011): **Randomized algorithms for matrices and data**. Stanford University. 1-54.
[32] MAHONEY, M. W. and DRINEAS, P. (2009): CUR matrix decompositions for improved data analysis, **PNAS**, 106, 697-702.
[33] MIRANDA, J., and BRINGAS, R. (2008): Analysis of DNA microarray data. Part I: Technological background and experimental design. **Biotecnología Aplicada**, 25, 90-96.
[34] PARMIGIANI, G., GARRETT, E. S., IRIZARRY, R. A., and ZEGER, S. L. (2003): The Analysis of Gene Expression Data: An Overview of Methods and Software. **The Analysis of Gene Expression Data**, 1-45. Springer Science & Business Media, Chichester.
[35] PAPAILIOPOULOS, D., KYRILLIDIS, A. and BOUTSIDIS, C. (2014): Provable Deterministic Leverage Score Sampling. In **Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining**, 997-1006.
[36] PENG, H., LONG, F. and DING, C. (2005): Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy, **IEEE Trans. Pattern Anal. Mach. Intell**, 27, 1226–1238.
[37] R CORE TEAM (2017): R: A language and environment for statistical computing. **R Foundation for Statistical Computing**, Vienna, Austria. URL https://www.R-project.org/.
[38] SAEYS, Y., INZA, I. AND LARRAÑAGA, P. (2007): A review of feature selection techniques in bioinformatics, **Bioinformatics**, 23, 2507–2517.
[39] SÁNCHEZ-MARROÑO, N., ALONSO-BETANZOS, A., and TOMBILLA-SANROMÁN, M. (2007): Filter methods for feature selection-a comparative study. In **International Conference on Intelligent Data Engineering and Automated Learning**, 178-187. Springer, Berlin, Heidelberg.
[40] VAN DER MAATEN, L., POSTMA, E., and VAN DEN HERIK, J. (2009): Dimensionality reduction: a review comparative. **J. Mach Learn Res**, 10, 1-35.
[41] VENABLES, W. N. and RIPLEY, B. D. (2002): **Modern Applied Statistics with S.** Fourth Edition. Springer, New York. ISBN 0-387-95457-0.
[42] Wang, N. N. (2009): **Statistical Problems in DNA Microarray Data Analysis** (Doctoral dissertation, UC Berkeley).
[43] WHEELER, D. A., SRINIVASAN, M., EGHOLM, M., SHEN, Y., CHEN, L., MCGUIRE, A., ... and ROTHBERG, J. M. (2008): The complete genome of an individual by massively parallel DNA sequencing. **Nature**, 452, 872-876.
[44] YU, L. and LIU, H. (2003). Feature selection for high-dimensional data: a fast correlation-based filter solution, In **Proceedings, Twentieth International Conference on Machine Learning**, 856-863.

APPENDIX I: Results of the loadings vectors for the first two principal components.

Groups	Genes	Component 1	Component 2	Groups	Genes	Component 1	Component 2
1	197	-0.1631	-0.0542	2	256	-0.1931	0.0145
1	857	-0.1873	-0.0546	2	1967	-0.0092	0.2236
1	498	-0.1747	-0.0467	1	476	-0.1208	-0.0684
1	486	-0.1900	-0.0272	2	1854	-0.1796	0.0366
1	1638	-0.1796	-0.0214	1	549	-0.0933	-0.0378
1	1496	-0.1667	-0.0328	1	1844	-0.1803	-0.0341
1	479	-0.1817	-0.0047	1	1157	-0.1190	-0.0548
2	897	-0.0129	0.2424	1	1463	-0.1285	-0.0142

2	1793	-0.1827	0.0105	1	1153	-0.0595	-0.0498
1	854	-0.1177	-0.0950	2	822	-0.0165	0.2420
1	1036	-0.1797	-0.0520	2	824	-0.0397	0.2165
1	530	-0.1799	-0.0218	2	36	-0.0945	0.0163
1	1397	-0.1906	-0.0042	1	75	-0.1382	-0.0706
2	1635	-0.052-6	0.2334	1	1094	-0.1606	-0.0119
1	1685	-0.1789	-0.0297	2	415	-0.0307	0.2215
1	625	-0.0860	-0.0503	1	258	-0.0561	-0.0482
1	1299	-0.1775	-0.0073	1	1461	-0.1815	-0.0163
1	1829	-0.1907	-0.0025	1	235	-0.0833	-0.0248
1	992	-0.0769	-0.0580	1	3	-0.0891	-0.0117
1	1974	-0.0243	0.2384	1	33	-0.0644	-0.0015
2	1274	-0.1733	0.0103	2	1387	-0.0763	0.1992
2	1494	-0.0428	0.2459	2	1421	-0.0546	0.2356
1	1972	-0.0621	-0.0200	2	323	-0.0601	0.2017
1	1478	-0.1779	-0.0406	2	929	-0.0200	0.2232
2	1247	-0.0464	0.2368	1	713	-0.0924	-0.0368
1	921	-0.1872	-0.0126	1	1077	-0.1318	-0.0847
1	443	-0.1927	-0.0023	2	249	-0.0231	0.2090
2	1843	-0.0540	0.2281	2	737	-0.0386	0.2288
2	806	-0.0227	0.2384	1	733	-0.0664	-0.0349
1	1384	-0.1761	-0.0004	1	295	-0.1051	-0.0287
2	993	-0.0456	0.2294	1	38	-0.0898	-0.0065