RESTRICTED CUR MATRIX DECOMPOSITION IN CANCER DNA MICROARRAY DATA CLASSIFICATION PROBLEMS.

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

ABSTRACT
DNA microarray data for cancer are datasets that originate from the use of DNA microarray technology in the classification of cancer tumors. These sets constitute a very complex classification problem for supervised and unsupervised data modeling. Due to their high dimensions in the number of columns, DNA microarray data in cancer constitute a Column Subset Selection Problem. This problem is closely related to the restricted CUR matrix decomposition. In this work we propose the restricted CUR matrix decomposition as a multivariate filter method. To do this, we consider the Frobenius norm relative error $\theta = \frac{\|A-C\cdot X\|_F}{\|A\|_F}$ as a finite succession in function of the number of columns selected for $n \times p$ with $n \ll p$ matrix A. Based on this result, we propose two algorithms that select a columns subset in such a way that $\theta = \frac{\|A-C\cdot X\|_F}{\|A\|_F}$ is as small as possible. We applied the proposed algorithms to six cancer DNA microarray datasets. The subsets selected by these algorithms are used as predictors to train the C4.5, NB and SVM classifiers, respectively. Using the 5-field cross-validation resampling technique, accuracy measure is calculated for these three classifiers. Finally, the results obtained are compared with the results found in the literature using Friedman's non-parametric test, concluding that these results are similar.

KEYWORDS: DNA microarray data, columns subset selection problems, multivariate filter method, restricted CUR matrix decomposition, Frobenius norm relative error.

MSC: 62H25

RESUMEN
Los datos de microarreglos de ADN para el cáncer son conjuntos que se originan por la utilización de la tecnología de microarreglos de ADN en la clasificación de tumores cancerígenos. Estos conjuntos constituyen un problema de clasificación muy complejo ante la modelación supervisada y no supervisada de datos. Debido a su alta dimensión en el número de columnas, los datos de microarreglos de ADN para el cáncer constituyen un Problema de Selección de Subconjunto de Columnas. Este problema está estrechamente relacionado con la descomposición matricial CUR restringida. En este trabajo proponemos la descomposición matricial CUR restringida como un método de filtro multivariado. Para ello, consideramos el error relativo de la norma de Frobenius $\theta = \frac{\|A-C\cdot X\|_F}{\|A\|_F}$ como una sucesión finita en función del número de columnas seleccionadas para una matriz A de orden $n \times p$ con $n \ll p$. A partir de este resultado proponemos dos algoritmos que buscan seleccionar un subconjunto de columnas de manera que el error relativo de la norma de Frobenius sea lo más pequeño posible. Aplicamos los algoritmos propuestos a seis conjuntos de datos de microarreglos de ADN para el cáncer. Los subconjuntos seleccionados por estos algoritmos son empleados como predictores para entrenar los clasificadores C4.5, NB y SVM, respectivamente. Usando la técnica de remuestreo por validación cruzada 5 campos, se calcula la medida de exactitud en estos tres clasificadores. Finalmente, los resultados obtenidos son comparados con los resultados que se encuentran en la literatura utilizando la prueba no paramétrica de Friedman, concluyendo que estos resultados son similares.

PALABRAS CLAVES: datos de microarreglos de ADN, problema de selección de subconjunto de columnas, método de filtro multivariado, descomposición matricial CUR restringida, error relativo de la norma de Frobenius.

1. INTRODUCTION

DNA microarray technology traces its origins to the Southern Blot (Southern, E. M., 1975), a foundational technique in molecular biology, and enables the simultaneous assessment of thousands of gene expressions on a solid surface (Heller, M. J., 2002). This analytical process comprises three key phases: (i) microarray fabrication; (ii) sample processing; and (iii) data interpretation (Ramírez-Salcedo, J. ***et al.***, 2014).

Despite their historical significance in functional genomics, DNA microarrays face persistent reproducibility challenges that limit their reliability in clinical and research settings (Tarca, A. L. ***et al.***, 2006). Reproducibility varies depending on the technological platform and experimental conditions. For technical replicates, oligonucleotide-based platforms such as Affymetrix and Agilent demonstrate high consistency, with correlation coefficients exceeding 0.9 (Bakay, M. ***et al.***, 2002). However, when comparing results across different versions of the same platform, such as Affymetrix HG95Av2 vs.

[1] yunier@uclv.cu

Affymetrix HG133, studies reveal significantly reduced correlations due to differences in probe sets targeting the same genes (Bammler, T. ***et al.***, 2005). cDNA microarrays exhibit even greater technical variability, with correlation coefficients ranging from 0.5 to 0.95, largely attributed to probe-printing inconsistencies (Jenssen, T. K. ***et al.***, 2002).

Cross-platform reproducibility remains a critical challenge, as highlighted by studies such as (Bammler, T. ***et al.,*** 2005; Jarvinen, A. K. ***et al.***, 2004; Carter, S. L. ***et al.***, 2005). A primary issue involves low-expression genes: Affymetrix and cDNA microarrays show poor concordance for these genes (Carter, S. L. ***et al.***, 2005). To address this, studies like Draghici, S. ***et al.*** (2006) excluded low-expression genes, improving cross-platform correlations. A second challenge concerns differentially expressed genes: platforms share only 10–30% of such genes, as demonstrated by Bammler, T. ***et al.*** (2005).

Applications of DNA microarrays include gene discovery, disease diagnosis, drug development, and toxicological research (Bednár M., 2000). In oncology, these tools generate datasets (DNA microarray data) for tumor classification (Heller, M. J., 2002), posing a complex challenge in supervised and unsupervised learning models (Wang, N. N., 2009). This complexity arises from the imbalance between the low sample size (tens to hundreds) and high dimensionality (thousands to tens of thousands of gene expression levels), a phenomenon termed the "large p small n problem" (Johnstone, I. and Titterington, D., 2009; Bolón-Canedo, V. ***et al.***, 2014).

The presence of the "large p small n" problem in these datasets makes DNA microarray data for cancer a Feature Subset Selection Problem (FSSP) (Inza, I. ***et al.***, 2004). For that reason, the search for the feature subset is performed through feature selection methods (Hambali, M. A. ***et al.***, 2020). These methods seek to obtain a simple mathematical model, easy to interpret and with high prediction in classification. To this end, feature selection methods work by eliminating from the selection process those features that are irrelevant and redundant (Li, J. ***et al.***, 2017).

On the other hand, the "large p small n" problem brings with it a high correlation among the levels of gene expressions (Gui, J. ***et al.***, 2010). This redundancy in gene expression levels makes it difficult to obtain a highly predictive and generalizable classification model (Boulesteix, A. L. and Strimmer, K., 2007). In that sense, feature extraction methods play a fundamental role in dimension reduction (Jolliffe, I. T., 2002). These methods seek to transform data from a high-dimensional space to a low-dimensional space by constructing new variables (Carreira-Perpinán, M. A., 1997). These new variables are called latent variables. Based on the construction of the latent variables, the feature extraction methods are classified as linear and non-linear, and depending on the use of the response variable, supervised and unsupervised (Hira, Z. M. and Gillies, D. F., 2015).

Deep learning emerges as a promising solution for high-dimensional problems in this domain (Hira, Z. M. and Gillies, D. F., 2015; Cleofas-Sánchez, L. ***et al.***, 2019; Geman, O. ***et al.***, 2016). Architectures such as recurrent neural networks (RNNs) (Medsker, L. R. and Jain, L., 2001) have demonstrated the ability to extract non-linear hierarchical patterns, outperforming traditional methods like SVM or KNN in accuracy (LeCun, Y. ***et al.***, 2015). For instance, Chen, D. ***et al.*** (2017) implemented an innovative approach using RNNs to differentiate between benign and malignant breast tumors. Their methodology integrated four recurrent neural networks dedicated to extracting clinical patterns, culminating in a fifth RNN specialized for the classification stage. Similarly, Abdel-Zaher, A. M. and Eldeib, A. M. (2016) designed an automated early breast cancer detection system, combining deep belief networks (DBNs) for initial training and backpropagation neural networks to optimize results.

In the context of genomic datasets with class imbalance, Reyes-Nava, A. ***et al.*** (2019) proposed a deep learning-based multi-layer perceptron (DL-MLP), specifically adapted to address heterogeneity in gene expression microarrays. Complementing these advances, Viloria, A. ***et al.*** (2020) evaluated the performance of MLP classifiers in complex environments characterized by high dimensionality, limited samples, and imbalanced distributions, demonstrating their applicability in precision genomics studies. For further details, it is recommended to consult (Hambali, M. A. ***et al.***, 2020).

Another way to reduce the dimension in these data sets is low-rank matrix approximation methods (Kishore Kumar, N. and Schneider, J., 2017). These methods seek to approximate a matrix A of order $m \times n$ by another matrix of lower rank $k \ll min\{m, n\}$ in order to obtain a more compact representation of the data with the least loss of information (Drineas, P. and Mahoney, M.W., 2016). The Singular Value Decomposition (SVD) is the most common for these purposes, since it provides the best approximation of rank k to A (Mahoney, M. W., 2011). However, this decomposition projects the matrix A onto the first k left or right singular vectors; which makes these vectors very difficult to interpret as they are linear combinations of the data (Kuruvilla, F. G. ***et al.***, 2002). Given this situation, more efficient low-rank matrix approximation techniques have been developed, such as CUR matrix decomposition (Mahoney, M. W. and Drineas, P., 2009).

The CUR matrix decomposition allows obtaining low-rank matrix approximations for a matrix A through the product of three matrices C, U and R. The matrices C and R contain some columns and rows of A,

respectively; while U is a matrix that is carefully constructed in a way that guarantees this approximation (Benjamin Erichson, N. ***et al***., 2018). To this end, this decomposition expresses the data in terms of a small number of columns and/or rows, which makes it easier to interpret (Mahoney, M. W. and Drineas, P., 2009). However, in the context of DNA microarray data for cancer, CUR matrix decomposition cannot be used because this kind of data constitutes a Column Subset Selection Problem (CSSP) (Mahoney, M. W. and Drineas, P., 2009; Bodor, A. ***et al.***, 2012; Rodríguez, Y. E. T., 2023). In this context, the subset of columns is the matrix C. This matrix is formed by the selection of k columns of a $m \times n$ matrix A of order such that the residual $\|A - P_C A\|_F$ is minimum for all the possible choices $\binom{n}{k}$ of C. Where $P_C = \mathrm{CC}^+$ denotes the projection onto the k-dimensional space generated by the columns of C, C^+ is the Moore-Penrose pseudoinverse of the matrix C and $\|\cdot\|_F$ denotes the Frobenius norm (Boutsidis, C., 2011).

This CSSP is closely related to the restricted CUR matrix decomposition (Boutsidis, C. ***et al.***, 2008b, Boutsidis, C. ***et al.***, 2009). This decomposition allows the matrix A to be approximated by the product of two matrices C and X, where C contains some columns of the matrix A, and $X = C^+ \cdot A$ so that the residual $\|A - C \cdot X\|_F$ is minimum. In this case the matrix C forms the subset of columns and $P_C A = C \cdot X$ (Mahoney, M. W. and Drineas, P., 2009).

Different types of restricted CUR matrix decompositions have been proposed in the scientific literature (Drineas, P. ***et al.***, 2006a; Drineas, P. ***et al.***, 2006b; Boutsidis, C. ***et al.***, 2008; Drineas, P ***et al.***, 2008; Boutsidis, C. ***et al.***, 2009; Mahoney, M. W. and Drineas, P., 2009; Boutsidis, C., 2011; Papailiopoulos, D. ***et al.***, 2014; Yang, J. ***et al***., 2015). These decompositions differ in the criteria for choosing the columns that form the matrix C and in the error rates obtained (Boutsidis, C. ***et al.***, 2008a; Boutsidis, C ***et al.***, 2008b; Boutsidis, C. ***et al.***, 2009; Avron, H. and Boutsidis, C., 2013; Papailiopoulos, D. ***et al.***, 2014; Boutsidis, C. and Woodruff, D. P., 2014; Belhadji, A. ***et al.***, 2020).

In the framework of DNA microarray data for cancer, this decomposition has been applied in:

- In 2008 Boutsidis ***et al.*** published "Unsupervised Feature Selection for Principal Components Analysis" (Boutsidis, C. ***et al.***, 2008b). The authors presented a restricted CUR matrix decomposition whose criterion for choosing the columns that form the matrix C is through a novel hybrid algorithm that works in two stages. From a theoretical point of view, their algorithm significantly improved the exact selection of k columns over existing algorithms for small to moderate values of k. They evaluated this algorithm in the area of genetics, in which they analyzed a matrix of order 90×2000000 that corresponded to 90 individuals from two Asian populations and 2000000 Simple Nucleotide Polymorphisms (SNPs) (The International HapMap Consortium, 2005). This dataset, derived from cDNA microarray technology, was also analyzed in (Drineas, P. ***et al.***, 2008).
- In Drineas, P. ***et al.***, 2008, the authors propose a restricted CUR matrix decomposition whose criterion for choosing the columns that form the matrix C follows an empirical sampling distribution. To build the matrix C, they created an algorithm that takes as input any $m \times n$ matrix A, a rank parameter k and an error parameter ϵ. The matrix X is calculated by the matrix product between $C^+$and A, with C^+ being the Moore-Penrose pseudoinverse of the matrix C. They comment that the algorithm to calculate the matrix C is the first to have polynomial time with theoretical guarantees on the approximation error.
- In 2009, Mahoney and Drineas (2009) proposed a restricted CUR matrix decomposition whose criterion for choosing the columns that form the matrix C consists of a normalized importance factor. This importance factor is 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 the 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 matrix X is calculated by the matrix product between $C^+$and A, with C^+ being the Moore-Penrose pseudoinverse of the matrix C. Finally, they applied this algorithm to a DNA microarray dataset of soft tissue tumors. This set contained 5520 rows and 31 columns represented by genes and samples, respectively. The columns were divided by three groups, represented by gastrointestinal stromal tumor (GIST), leiomyosarcoma (LEIO) and synovial sarcoma (SARC).
- In 2012, the restricted CUR matrix decomposition proposed by Mahoney and Drineas (2009) is implemented in the rCUR package (Bodor, A. and Solymosi, N., 2012) of the R software. In this package the ColumnSelect algorithm is programmed and four more variants derived from it. In Bodor, A. ***et al***., 2012, the authors show the use of the rCUR package in feature selection choosing the first 27 genes with highest importance factors from the DNA microarray dataset of soft tissue tumors (Nielsen, T. ***et al.***, 2002).
- In 2023, the restricted CUR matrix decomposition is proposed as a gene subset selection method in DNA microarray data for cancer (Rodríguez, Y. E. T., 2023). This method seeks to minimize the

normalized Frobenius norm error $\theta = \frac{\|A - C \cdot X\|_F}{\|A - A_k\|_F}$ by approximating the data matrix by a low-rank matrix. Based on this result, the author proposed an algorithm that seeks to select a subset of genes in a non-conventional way to filter methods. Finally, the algorithm is applied to a Colon cancer DNA microarray dataset. This set contains the expression levels of 2000 genes for 62 patients divided into two classes: the sick class and the healthy class.

In this work we propose the restricted CUR matrix decomposition as a multivariate filter method. To do this, we consider the Frobenius norm relative error $\theta = \frac{\|A - C \cdot X\|_F}{\|A\|_F}$ as a finite succession in function of the number of columns selected for a $n \times p$ with $n \ll p$ matrix A. We show that this error is an decreasing succession and reaches its minimum faster than the normalized Frobenius norm error $\theta = \frac{\|A - C \cdot X\|_F}{\|A - A_k\|_F}$. Based on this result, we propose two algorithms that select a columns subset in such a way that $\theta = \frac{\|A - C \cdot X\|_F}{\|A\|_F}$ is as small as possible. We applied the proposed algorithms to six cancer DNA microarray datasets. The subsets selected by these algorithms are used as predictors to train the C4.5, NB and SVM classifiers, respectively. Using the 5-field cross-validation resampling technique, balanced accuracy measures are calculated for these three classifiers. Finally, the results obtained are compared with the results found in the literature using non-parametric hypothesis tests in paired samples, concluding that these results are similar.

The paper is organized as follows: Section 2 defines the Frobenius norm relative error $\theta = \frac{\|A - C \cdot X\|_F}{\|A\|_F}$ as a finite succession in function of the number of columns selected for $n \times p$ with $n \ll p$ matrix A. Section 3 proposes the restricted CUR matrix decomposition as a multivariate filter method. Section 4 details the proposed algorithms. Section 5 presents the research datasets. Sections 6 and 7 show the results and discussion of the work, respectively. Finally, Section 8 presents the conclusions of the document.

2. FROBENIUS NORM RELATIVE ERROR

In 2023, Rodríguez, Y. E. T (2023), proposed the Frobenius norm relative error $\theta = \frac{\|A - C \cdot X\|_F}{\|A\|_F}$ as a finite succession in function of the number of columns selected for $n \times p$ with $n \ll p$ matrix A. The author demonstrated that this succession is decreasing and tends to zero.

Similarly, the Frobenius norm relative error $\frac{\|A - P_C A\|_F}{\|A\|_F}$ can be considered as a function of the number of selected columns, this is: for all finite natural number p and for each c such that $c \ll p$, one can define $\theta_R(c) = \frac{\|A - C(c) \cdot X\|_F}{\|A\|_F}$ and form the finite succession $\{\theta_R(c)\}_{c=1}^{p}$. This succession reaches its minimum value faster than the succession formed by $\{\theta_N(c)\}_{c=1}^{p}$ with $\theta_N(c) = \frac{\|A - C(c) \cdot X\|_F}{\|A - A_k\|_F}$.

In order to prove this, we will first show that the succession $\{\theta_R(c)\}_{c=1}^{p}$ is decreasing. Let us consider consider $P_C A = C \cdot X$, the projection onto the k-dimensional generated space by the columns of C in the matrix A. For each $c = 1, \ldots, p$, let $C(c)$ denote the matrix consisting of the first c columns selected by the restricted CUR matrix decomposition, respectively.

For any c such that $1 \leq c \leq p-1$, consider $C(c)$ and $C(c+1)$. By construction, $C(c+1)$ contains all columns of $C(c)$ plus one additional column. Consequently, the column space of $C(c)$ is contained in the column space of $C(c+1)$; that is, $R(C(c)) \subseteq R(C(c+1))$.

The orthogonal projection $P_{C(c)}A$ is the best approximation to A in $R(C(c))$ with respect to the Frobenius norm; in other words, it minimizes $\|A - B\|_F$ over all matrices B in $R(C(c))$. Similarly, $P_{C(c+1)}A$ is the best approximation to A in $R(C(c+1))$.

Since $R(C(c)) \subseteq R(C(c+1))$ the set of admissible approximations in $R(C(c+1))$ includes every approximation available in $R(C(c))$. Therefore, the approximation error over $R(C(c+1))$ cannot exceed that over $R(C(c))$. Hence, $\left\|A - P_{C(c+1)}A\right\|_F \leq \left\|A - P_{C(c)}A\right\|_F$.

Thus, $\theta_R(c) > \theta_R(c+1)\ \forall c \geq 1$.

In order to prove that the succession $\{\theta_R(c)\}_{c=1}^{p}$ reaches its minimum value faster than the succession $\{\theta_N(c)\}_{c=1}^{p}$, we will assume that A is a full-rank matrix, this is, $r = min(n, p) = n$.

$\|A\|_F = \sqrt{\sum_{i=1}^{n} \sum_{j=1}^{p} (a_{ij})^2}$ (Definition of Frobenius norm)

$= \sqrt{\sum_{i=1}^{n} (\sigma_i)^2}$ (Frobenius norm property: σ_i is the i-th singular value of A)

Therefore, $\|A\|_F^2 = \sum_{i=1}^{n} (\sigma_i)^2$.

$\|A\|_F^2 = \sum_{i=1}^{k} (\sigma_i)^2 + \sum_{i=k+1}^{n} (\sigma_i)^2$ (Numerical series property)

$= \|A_k\|_F^2 + \sum_{i=k+1}^{n}(\sigma_i)^2$ (Definition of Frobenius norm for a rank-k matrix)

$= \|A_k\|_F^2 + \|A - A_k\|_F^2$ (Eckart-Young approximation theorem)

Finally, $\|A\|_F^2 > \|A - A_k\|_F^2$, therefore, $\|A\|_F > \|A - A_k\|_F$.

Then, $\frac{1}{\|A\|_F} < \frac{1}{\|A-A_k\|_F}$ and multiplying both members of the inequality by $\left\|A - P_{C(c)}A\right\|$ we get $\frac{\|A-P_{C(c)}A\|}{\|A\|_F} < \frac{\|A-P_{C(c)}A\|}{\|A-A_k\|_F}$.

On the other hand, $\left\|A - P_{C(c)}A\right\| > 0$ when $c < n$ and $\left\|A - P_{C(c)}A\right\| = 0$ when $c \geq n$ since A is a full-rank matrix.

Then, $0 < \theta_R(c) < \theta_N(c)\ \forall c < n$ and $\theta_R(c) = \theta_N(c) = 0\ \forall c \geq n$.

Therefore, the finites successions $\{\theta_R(c)\}$ and $\{\theta_N(c)\}$ reaches their minimum value when $c = n$, however, $\{\theta_R(c)\}$ does so faster due to $\theta_R(c) < \theta_N(c)$.

3. RESTRICTED CUR MATRIX DECOMPOSITION: A MULTIVARIATE FILTER METHOD

Let A be an $n \times p$ matrix with $n \ll p$, containing n observations with known class labels and p features. Let Y be an $n \times 1$ column vector encoding the class labels (0 or 1) of the n observations. It is assumed that: $E(A_j) = 0,\ Var(A_j) = 1,\ Cor(A_j,\ Y) \neq 0, \forall j \in \{1, \ldots, p\}$, where A_j is the j-th column of A.

Considering that the rows of matrix A remain fixed during the selection process of the restricted CUR matrix decomposition, it is valid to consider the Frobenius norm relative error $\frac{\|A-C\cdot X\|_F}{\|A\|_F}$ as a finite succession. Section 2 demonstrated that this succession is monotonically decreasing and reaches its minimum value for all $c \geq n$. In this context, the restricted CUR matrix decomposition can be considered as a columns subset selection method (Rodríguez, Y. E. T., 2023). In particular, as a filter method (Mahoney, M. W. and Drineas, P., 2009; Bodor, A. ***et al.***, 2012). In that sense, the subset of columns is constituted by the matrix C such that $\frac{\|A-C\cdot X\|_F}{\|A\|_F}$ is minimal. This selection process is performed without taking into account the classifier, a characteristic that defines filter methods.

From the above result, we propose a multivariate filter method that seeks to minimize the Frobenius norm relative error as small as possible. This method seeks to select a number $\mathcal{C} = c(\varepsilon)$ such that for $\mathcal{C} < n$ the inequality $\theta_R(\mathcal{C}) < \varepsilon$ is satisfied for any number $\varepsilon > 0$ (however small it may be). Where the matrix $C(\mathcal{C})$ is the subset of columns.

In this way, we can compute a classification model between the columns of matrix $C(\mathcal{C})$ and the column vector Y. Then, using the k-field cross-validation resampling technique for validates the model.

4. PROPOSED ALGORITHMS

4.1. rCURd ALGORITHM

The **r**estricted **CUR d**ecomposition (**rCURd)** seeks to select a subset of columns by minimizing the Frobenius norm relative error by approximating the data matrix A by a low rank matrix C. This algorithm uses the restricted CUR matrix decomposition proposed by Mahoney and Drineas, whose criterion for choosing the columns that form the matrix C consists of a normalized importance factor (Mahoney and Drineas, 2009). This importance factor is defined by $\pi_j = \frac{1}{k}\sum_{i=1}^{k}\left(v_j^i\right)^2, \forall j = 1, \ldots, p$ where v_j^i is the j-th component of the i-th right singular vector of A. The rank parameter k is set by of the high variability explained by the principal components (top k right singular vectors).

The criterion used by rCURd consists of selecting c columns to form the matrix C, retaining the columns with the highest importance factors. This criterion enjoys great acceptance by the scientific community as evidenced by the research of (Mahoney, M. W. and Drineas, P., 2009; Bodor, A. ***et al.***, 2012; Rodríguez, Y. E. T. ***et al.***, 2012; Yang, J. ***et al.***, 2015; Barahona, G. V. ***et al.***, 2019). In the rCUR package (Bodor, A. and Solymosi, N., 2012) it is implemented as *top.scores*. A description of rCURd is given in Table 1.

Input: Matrix $A_{n\times p}$ with $n \ll p$, rank parameter k, $0 < \varepsilon << 1$
Output: Subset of columns selected $C_{n\times c}$ with $c \ll p$.
1: Compute the normalized importance factors $\pi_j = \frac{1}{k}\sum_{i=1}^{k}\left(v_j^i\right)^2$ of A

2: **for** c = 1 **to** p **do**
3: Compute the matrix $C(c)$ retaining the columns with the highest importance factors π_j of A
4: Compute the pseudo-inverse of matrix $C(c)$, denoted by $C^+(c)$
5: Compute the matrix X by the matrix product between $C^+(c)$ y A
6: Compute the Frobenius norm relative error $\theta_R(c) = \frac{\|A - C(c) \cdot X\|_F}{\|A\|_F}$
7: **if** $\theta_R(c) < \varepsilon$ **then** {
8: Select the columns subset of A by the matrix $C = C(\mathrm{c})$
9: **break**
10: }
11: **end**

Table 1: Description of the rCURd algorithm.

4.2 FoS-rCURd ALGORITHM

In this section, we propose an algorithm that presents a significant improvement in execution time compared to the rCURd algorithm. This algorithm, called **Fo**rward **S**election by **r**estricted **CUR d**ecomposition (**FoS-rCURd**), seeks to select a subset of columns by obtaining a sample that is as representative as possible of the population. To do this, FoS-rCURd works in two steps. In the first step, a sample consisting of the h first terms of the finite succession $\{\theta_R(c)\}_{c=1}^{p}$ considered as the initial population is selected. Subsequently, it is analyzed how representative the sample is with respect to the population. If it is not representative, the h first terms are deleted and the succession $\{\theta_R(c)\}_{c=h+1}^{p}$ is considered a new population and the h first terms of it are considered a new sample. This iterative process is carried out until the sample is as representative of the population as possible. In the second step, the c-th terms are computed: $\theta_R(c) = \frac{\|A - C(c) \cdot X\|_F}{\|A\|_F}$ in the selected sample. Finally, the subset of genes selected by FoS-rCURd is $C = \min_{c_1 \le c \le c_{1+h}} \{C(c) : \theta_R(c) < \varepsilon\}$ with c_1 being the first element of this sample. A description of the FoS-rCURd algorithm is given in Table 2.

Input: Matrix $A_{n \times p}$ with $n \ll p$, rank parameter k, $h \ll n$, $0 < \varepsilon << 1$
Output: Subset of columns selected $C_{n \times c}$ with $c \ll p$.
1: Compute the normalized importance factors π_j of A
2: **for** i = 1 **to** p **do**
3: Compute the base of the i-th rectangle $b = (i - (i-1)) \cdot h$
4: Compute the matrix $C(i \cdot h)$ retaining the columns with the highest importance factors π_j of A
5: Compute the pseudo-inverse of matrix $C(i \cdot h)$, denoted by $C^+(i \cdot h)$
6: Compute the matrix X by the matrix product between $C^+(i \cdot h)$ and A
7: Compute the height of the i-th rectangle $a = \frac{\|A - C(i \cdot h) \cdot X\|_F}{\|A\|_F}$
8: **if** $a \cdot b < \varepsilon \cdot b$ **then** {
9: **for** c = $(i-1) \cdot h$ **to** $i \cdot h$ **do**
10: Compute the matrix $C(\mathrm{c})$ retaining the columns with the highest importance factors π_j of A
11: Compute the pseudo-inverse of matrix $C(\mathrm{c})$, denoted by $C^+(\mathrm{c})$
12: Compute the matrix X by the matrix product between $C^+(\mathrm{c})$ and A
13: Compute the Frobenius norm relative error $\theta_R(\mathrm{c}) = \frac{\|A - C(\mathrm{c}) \cdot X\|_F}{\|A\|_F}$
14: **if** $\theta_R(\mathrm{c}) < \varepsilon$ **then** {
15: **break**
16: }
17: Select the columns subset of A by the matrix $C = C(\mathrm{c})$
18: **end**
19: **break**
20: }
21: **end**

Table 2: Description of the FoS-rCURd algorithm.

5. DATASET

We selected six cancer DNA microarray datasets from the research "A review of microarray datasets and applied feature selection methods" by Bolón-Canedo, V. ***et al.*** (2014). These sets were selected on the following objective: Solve the "large p small n" problem in the datasets through the subsets of genes selected by the proposed algorithms. Table 3 provides a summary of these dataset. S denote the number of

samples, F represents the number of feature, IR[2] indicates the imbalance ratio, and the last two columns correspond to the original reference of the dataset and its source URL.

.

Dataset	S	F	IR	Original Ref.	Download
CNS	60	7129	1.86	(Pomeroy, S. ***et al.***, 2002)	(Broad institute. Cancer Program Data Sets, 2017)
Colon	62	2000	1.82	(Alon, U. ***et al.***, 1999)	(Kent Ridge Bio-Medical Dataset, 2017)
DLBCl	47	4026	1.04	(Alizadeh, A. ***et al.***, 2000)	(Kent Ridge Bio-Medical Dataset, 2017)
GLI-85	85	22283	2.27	(Freije, W. ***et al.***, 2004)	(Feature Selection Datasets at Arizona State University, 2017)
Ovarian	253	15154	1.78	(Petricoin, E. ***et al.***, 2002)	(Kent Ridge Bio-Medical Dataset, 2017)
Smk	136	12600	1.34	(Spira, A. ***et al.***, 2007)	(Feature Selection Datasets at Arizona State University, 2017)

Tabla 3: Cancer DNA microarray datasets description

6. RESULTS

The research results were obtained using the R-3.4.3 software (R Core Team, 2017) on a computer with Windows 10 Pro 64-bit, 8 GB of RAM, an Inter(R) Core (TM) i3- 6100 and 1000 GB capacity.
To fulfill the abovementioned objective, a preprocessing step was applied to these dataset. This preprocessing involved centering and scaling the data, as well as selecting columns correlated with the dependent variable, where a value of 0 indicates a cancerous sample and a value of 1 indicates a non-cancerous sample, respectively. The following functions from the stats package (R Core Team, 2023) were used:

- The *scale* function to center the data (zero mean) and scale the data (unit variance) when necessary.
- The *cor.test* function to identify columns correlated with the dependent variable. This function implements a statistical test based on Pearson's product moment correlation coefficient. For this test, the level of significance was considered $\alpha = 0.05$.

Table 4 summarizes the preprocessing steps applied to these dataset. S denotes the number of samples, F represents the number of feature, IR indicates the imbalance ratio, and the last two columns corresponds the centering and scaling processes applied to the data.

.

Dataset	S	F	IR	Center	Scale
CNS	60	334	1.86	Yes	Yes
Colon	62	389	1.82	Yes	Yes
DLBCl	47	946	1.04	Yes	Yes
GLI-85	85	7057	2.27	Yes	Yes
Ovarian	253	8393	1.78	Yes	No
Smk	136	4919	1.34	Yes	Yes

Table 4: Cancer DNA microarray datasets description after preprocessing

After the data preprocessing step, an empirical study was carried out to determine the most appropriate threshold (ε) in selecting the number of columns of the matrix C (selected subset of genes). To do this, different thresholds were tested in the infinitely small succession in order to evaluate the impact of each threshold on the quality of the approximation. We begin with a threshold equal to 10%, and then gradually decrease with a difference of one unit to a threshold equal to 1%. This study was implemented in the RStudio integrated development environment. The packages used in the implementation were Matrix (Douglas, B. and Martin, M., 2017), MASS (Venables, W. N. and Ripley, B. D., 2002), rCUR, forach (Revolution Analytics and Weston, S., 2015) and ggplot2 (Wickham, H., 2009).
The results achieved in the empirical study for the selection of gene subsets in the DNA microarray data after preprocessing (see table 4), were computed with the rCURd and FoS-rCURd algorithms:

- CNS microarray data, we worked with the rCURd algorithm with rank parameter k equal to 21 and with the FoS-rCURd algorithm with rank parameter k and sample size h equal to 21 and 15, respectively.

[2] This ratio is defined as the number of negative class samples divided by the number of positive class sample, in which a high level indicates that the dataset is highly imbalanced.

- Colon microarray data, we worked with the rCURd algorithm with rank parameter k equal to 8 and with the FoS-rCURd algorithm with rank parameter k and sample size h equal to 8 and 15, respectively.
- DLBCL microarray data, we worked with the rCURd algorithm with rank parameter k equal to 21 and with the FoS-rCURd algorithm with rank parameter k and sample size h equal to 21 and 15, respectively.
- GLI85 microarray data, we worked with the rCURd algorithm with rank parameter k equal to 36 and with the FoS-rCURd algorithm with rank parameter k and sample size h equal to 36 and 15, respectively.
- Ovarian microarray data, we worked with the rCURd algorithm with rank parameter k equal to 1 and with the FoS-rCURd algorithm with rank parameter k and sample size h equal to 1 and 25, respectively.
- Smk microarray data, we worked with the rCURd algorithm with rank parameter k equal to 35 and with the FoS-rCURd algorithm with rank parameter k and sample size h equal to 35 and 19, respectively.

For each of the previous cases, the rank parameter k was set corresponding to 80% of the variance explained by the principal components. Table 5 shows these results.

.

Algorihm	Study	Cancer DNA microarray datasets					
		CNS	Colon	DLBCL	GLI85	Ovarian	Smk
rCURd	ε=0.10	58(0.0779)	59(0.0943)	**45(0.968)**	82 (0.0979)	182(0.1000)	179(0.0983)
	ε=0.09	58(0.0779)	60(0.0851)	46(0.0000)	83(0.0698)	196(0.0895)	181(0.0820)
	ε=0.08	**58(0.0779)**	61(0.0725)	46(0.0000)	83(0.0698)	207(0.0792)	182(0.0735)
	ε=0.07	59(0.0000)	62(0.0592)	46(0.0000)	**83(0.0698)**	217(0.0699)	183(0.0647)
	ε=0.06	59(0.0000)	62(0.0592)	46(0.0000)	84(0.0000)	227(0.0595)	184(0.0528)
	ε=0.05	59(0.0000)	**63(0.0423)**	46(0.0000)	84(0.0000)	234(0.0486)	185(0.0382)
	ε=0.04	59(0.0000)	64(0.0000)	46(0.0000)	84(0.0000)	240(0.0392)	**185(0.0382)**
	ε=0.03	59(0.0000)	64(0.0000)	46(0.0000)	84(0.0000)	247(0.0280)	186(0.0000)
	ε=0.02	59(0.0000)	64(0.0000)	46(0.0000)	84(0.0000)	**251(0.0173)**	186(0.0000)
	ε =0.01	59(0.0000)	64(0.0000)	46(0.0000)	84(0.0000)	253(0.0000)	186(0.0000)
FoS-rCURd	ε=0.10	58(0.0779)	59(0.0943)	**45(0.968)**	82(0.0979)	182(0.1000)	179(0.0983)
	ε=0.09	58(0.0779)	60(0.0851)	46(0.0000)	83(0.0698)	196(0.0895)	181(0.0820)
	ε=0.08	**58(0.0779)**	61(0.0725)	46(0.0000)	83(0.0698)	207(0.0792)	182(0.0735)
	ε=0.07	59(0.0000)	62(0.0592)	46(0.0000)	**83(0.0698)**	217(0.0699)	183(0.0647)
	ε=0.06	59(0.0000)	62(0.0592)	46(0.0000)	84(0.0000)	227(0.0595)	184(0.0528)
	ε=0.05	59(0.0000)	**63(0.0423)**	46(0.0000)	84(0.0000)	234(0.0486)	185(0.0382)
	ε=0.04	59(0.0000)	64(0.0000)	46(0.0000)	84(0.0000)	240(0.0392)	**185(0.0382)**
	ε=0.03	59(0.0000)	64(0.0000)	46(0.0000)	84(0.0000)	247(0.0280)	186(0.0000)
	ε=0.02	59(0.0000)	64(0.0000)	46(0.0000)	84(0.0000)	**251(0.0173)**	186(0.0000)
	ε=0.01	59(0.0000)	64(0.0000)	46(0.0000)	84(0.0000)	253(0.0000)	186(0.0000)

Table 5: Results of the empirical study for the selection of the subsets of genes in the DNA microarray data

After analyzing the results of table 5, the following conclusions were reached:

- From the CNS microarray data, the gene subset $C = C(58)$ was selected. This subset contains the expression levels of 58 genes for 60 patients with Medulloblastoma. To carry out this selection, we worked with an infinitely small succession generated by the rCURd and FoS-rCURd algorithms with threshold $\varepsilon = 0.08$. This threshold corresponds to the quality of the approximation, which is 0.0779 for a total of 58 levels of gene expressions.
- From the Colon microarray data, the $C = C(63)$ gene subset was selected. This subset contains the expression levels of 63 genes for 62 patients divided into two classes: the sick class and the healthy class. To carry out this selection, we worked with an infinitely small succession generated by the rCURd and FoS-rCURd algorithms with threshold $\varepsilon = 0.05$. This threshold corresponds to the quality of the approximation, which is 0.0423 for a total of 63 levels of gene expressions.
- From the DLBCL microarray data, the $C = C(45)$ gene subset was selected. This subset contains the expression levels of 45 genes for 47 patients with diffuse large B-cell lymphoma. To carry out this selection, we worked with an infinitely small succession generated by the rCURd and FoS-rCURd algorithms with threshold $\varepsilon = 0.10$. This threshold corresponds to the quality of the approximation, which is 0.0968 for a total of 45 levels of gene expressions.

- From the GLI85 microarray data, the $C = C(83)$ gene subset was selected. This subset contains the expression levels of 83 genes on 85 diffuse infiltrating gliomas. To carry out this selection, we worked with an infinitely small succession generated by the rCURd and FoS-rCURd algorithms with threshold $\varepsilon = 0.07$. This threshold corresponds to the quality of the approximation. which is 0.0698 for a total of 83 gene expression levels.
- From the Ovarian microarray data the $C = C(251)$ gene subset was selected. This subset contains the expression levels of 251 genes for 253 samples divided into two classes: the diseased class and the healthy class. To carry out this selection, we worked with an infinitely small succession generated by the rCURd and FoS-rCURd algorithms with threshold $\varepsilon = 0.02$. This threshold corresponds to the quality of the approximation, which is 0.0173 for a total of 251 gene expression levels.
- From the Smk microarray data, the $C = C(185)$ gene subset was selected. This subset contains the expression levels of 185 genes for 187 samples divided into two classes: the tumor class and the control class. To carry out this selection, we worked with an infinitely small succession generated by the rCURd and FoS-rCURd algorithms with threshold $\varepsilon = 0.04$. This threshold corresponds to the quality of the approximation, which is 0.0382 for a total of 185 gene expression levels.

It is not a surprise that both algorithms generated the same infinitely small succession; since the two algorithms seek to select a subset of genes so that the Frobenius norm relative error is as small as possible. However, the FoS-rCURd algorithm generated the infinitely small succession in much less time than the rCURd algorithm. This is because the FoS-rCURd algorithm presented a significant improvement in terms of its execution time. Table 6 shows these results.

Execution time	Cancer DNA microarray datasets					
	CNS	Colon	DLBCL	GLI85	Ovarian	Smk
rCURd (seconds)	0.64	0.61	0.64	16.80	200.77	54.2
FoS-rCURd (seconds)	0.26	0.19	0.20	4.83	33.17	11.64

Table 6: Results of the execution time for the selection of the subsets of genes in the DNA microarray data.

7. DISCUSSION

The results of the research showed that the FoS-rCURd algorithm was superior to the rCURd algorithm in solving the "large p small n" problem presented in the datasets studied. Based on this result, it was decided to study how good the FoS-rCURd algorithm was in selecting gene subsets.

Taking into account that the selection process of the FoS-rCURd algorithm was performed without considering the classifier, a review of the most commonly used filtering methods on cancer DNA microarray datasets was carried out. The results of the review yielded 2 articles that contained the six research datasets. These two articles were (Bolón-Canedo, V. ***et al.***, 2011) and (Bolón-Canedo, V. ***et al.***, 2014). However, the research by Bolón-Canedo, V. ***et al.*** (2014) was the most complete.

In (Bolón-Canedo, V. ***et al.***, 2014), the authors carried out an exhaustive review of the existing literature on DNA microarray datasets for cancer and feature selection methods applied in this context. In this sense, the authors used six feature selection methods by filters: Correlation-based Feature Selection (CFS) (Hall, M., 1999), Fast Correlation-Based Filter (FCBF) (Yu, L. and Liu, H., 2003), Interact (INT) (Zhao, Z. and Liu, H., 2007), Information Gain (IG) (Hall, M. and Smith, L., 1998), ReliefF (RF) (Kononenko, I., 1994) and minimun Redundancy Maximun Relevance (mRMR) (Peng, H. ***et al.***, 2005). The subsets selected by these methods were used as predictors in the C4.5 classification tree, the Naive Bayes (NB) classifier and the support vector machine with linear kernel (SVM), respectively. The measure of accuracy was calculated in order to validate the calculated classification models.

Following this idea, the subsets of genes selected by the FoS-rCURd algorithm were used as predictors for the C4.5, NB and SVM classification models, respectively. During the training process of these three classifiers, the 5-field cross-validation resampling technique was used. Finally, measures of accuracy were calculated in order to validate the calculated classification models. The packages used to obtain the results were RWeka (Hornik, K. ***et al.***, 2009), klaR (Weihs, C. ***et al.***, 2005), kernlab (Karatzoglou, A. ***et al.***, 2023) and caret (Kuhn, M., 2008).

In order to analyze the results achieved by the FoS-rCURd algorithm in problems of classification of cancer tumors, it was decided to perform the Friedman's non-parametric test. The hypothesis test was formulated as $\mathrm{H}_0: \mathrm{M}_{\mathrm{FoS-rCURd}} = \mathrm{M}_{\mathrm{CFS}} = \mathrm{M}_{\mathrm{FCBF}} = \mathrm{M}_{\mathrm{INT}} = \mathrm{M}_{\mathrm{IG}} = \mathrm{M}_{\mathrm{ReliefF}} = \mathrm{M}_{\mathrm{mRMR}} = \mathrm{M}_{\mathrm{Algorithms}}$ vs H_1: at least

the median of the balanced accuracy results in one algorithm was different from the others. In this hypothesis tests, the level of significance was considered $\alpha = 0.05$.
For each classifier, the p-value was higher than the significance level used. This means that there was no evidence in statistical terms to reject the null hypothesis. Therefore, we can conclude that the FoS-rCURd, CFS, FCBF, INT, IG, ReliefF and mRMR algorithms showed similar results (see table 7).

	Algorithms	DNA Microarray Datasets for Cancer						Friedman's test
		CNS	Colon	DLBCL	GLI85	Ovarian	Smk	
Classifier C4.5	FoS-rCURd	**0.83**	0.82	0.67	0.78	0.81	**0.76**	p-value = **0.4041**
	CFS	0.62	0.79	**0.75**	0.79	0.98	0.64	
	FCBF	0.48	0.79	0.73	**0.82**	**0.99**	0.61	
	INT	0.55	0.79	0.70	0.78	0.98	0.59	
	IG	0.63	**0.84**	0.73	0.81	0.96	0.65	
	ReliefF	0.53	0.82	0.73	**0.82**	**0.99**	0.61	
	mRMR	0.58	0.82	0.73	0.80	0.97	0.62	
Classifier NB	FoS-rCURd	**0.84**	0.81	**0.96**	**0.94**	0.67	**0.77**	p-value = **0.3234**
	CFS	0.67	**0.85**	0.90	0.82	**1.00**	0.65	
	FCBF	0.70	0.80	0.90	0.85	0.99	0.69	
	INT	0.70	0.77	0.90	0.82	**1.00**	0.64	
	IG	0.63	0.77	0.92	0.85	0.98	0.66	
	ReliefF	0.67	0.84	0.92	0.89	0.98	0.67	
	mRMR	0.62	0.80	0.94	0.80	0.99	0.67	
Classifier SVM	FoS-rCURd	**0.93**	0.74	**0.98**	**1.00**	0.98	0.70	p-value = **0.1488**
	CFS	0.62	0.81	0.88	0.88	**1.00**	0.64	
	FCBF	0.65	0.84	0.81	0.87	**1.00**	0.71	
	INT	0.62	0.81	0.88	0.88	**1.00**	0.66	
	IG	0.67	**0.85**	0.94	0.86	**1.00**	**0.72**	
	ReliefF	0.73	**0.85**	0.92	0.89	**1.00**	0.69	
	mRMR	0.70	0.84	0.96	0.89	**1.00**	0.68	

Table 7: Results of the Friedman's non-parametric test for balanced accuracy by the classifiers with the 5-field cross-validation resampling technique.

8. CONCLUSIONS

In this work, we propose the restricted CUR matrix decomposition as a multivariate filter method in a $n \times p$ with $n \ll p$ matrix A. This method tries to minimize the Frobenius norm relative error by approximating the data matrix by a low rank matrix. To do this, we considered the Frobenius norm relative error as a finite succession in function of the number of columns selected and it was shown that it is an decreasing succession. Based on this result, two algorithms were proposed for columns subset selection in a non-conventional way to filter methods. Both algorithms use the restricted CUR matrix decomposition proposed by Mahoney and Drineas, whose criterion for choosing the columns that form the matrix C consists of retaining those columns with the highest importance factors.
We applied the proposed algorithms, denoted by rCURd and FoS-rCURd, to six cancer DNA microarray dataset. To do this, an empirical study was carried out to determine the most appropriate threshold in selecting the number of columns of the matrix C. We started with a threshold equal to 10%, and then gradually decreased with a difference of one unit until a threshold equal to 1%. For each of the studies, the range parameter k was set corresponding to 80% of the variance explained by the principal components.
In all datasets the proposed algorithms selected the same subset of genes. However, the FoS-rCURd algorithm performed it in much less time than the rCURd algorithm. This is because the FoS-rCURd algorithm presented a significant improvement in terms of its runtime. Based on these results, the FoS-rCURd algorithm was superior to the rCURd algorithm in solving the "large p small n" problem presented in the datasets studied.
Finally, the subsets selected by the FoS-rCURd algorithm were used as predictors to train the C4.5, NB and SVM classifiers, respectively. Using the 5-field cross-validation resampling technique, accuracy measures were calculated for these three classifiers. The results obtained by these three classifiers were statistically compared with the results reported by the filter methods: CFS, FCBF, INT, IG, RF and mRMR. To do this, Friedman's non-parametric test was used for each classifier. The results of these hypothesis tests showed that the FoS-rCURd algorithm showed similar results to CFS, FCBF, INT, IG, RF and mRMR.

ACKNOWLEDGMENTS: The author wishes to thank the contributions of Dr. Valia Guerra Ones, Dr. Jesús Eladio Sánchez García, Dr. Uvedel Bernabé del Pino Paz and MSc. Lorenzo Antonio Pérez Carballo.

RECEIVED: APRIL, 2024.
REVISED: SEPTEMBER, 2025.

REFERENCES

[1] ABDEL-ZAHER, A. M. and ELDEIB, A. M. (2016): Breast cancer classification using deep belief networks. **Expert Systems with Applications**, 46, 139–144.
[2] ALIZADEH, A., EISEN, M., DAVIS, R., MA, C., LOSSOS, I., ROSENWALD, A., BOLDRICK, J., SABET, H., TRAN, T., YU, X. ***et al***. (2000): Distinct types of diffuse large b-cell lymphoma identified by gene expression profiling, **Nature**, 403(6769), 503–511.
[3] ALON, U., BARKAI, N., NOTTERMAN, D., GISH, K., YBARRA, S., MACK, D. and LEVINE, A. (1999): Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays, **Proc. Nat. Acad. Sci**, 96(12), 6745–6750.
[4] AVRON, H. and BOUTSIDIS, C. (2013): Faster subset selection for matrices and applications. **SIAM Journal on Matrix Analysis and Applications**, 34(4), 1464-1499.
[5] BAKAY, M., CHEN, Y. W., BORUP, R., ZHAO, P., NAGARAJU, K. and HOFFMAN, E. P. (2002): Sources of variability and effect of experimental approach on expression profiling data interpretation. **BMC Bioinformatics**, 3(4). [PubMed: 11936955]
[6] BAMMLER, T., BEYER, R. P., BHATTACHARYA, S., BOORMAN, G. A., BOYLES, A., BRADFORD, B.U., ***et al.*** (2005): Standardizing global gene expression analysis between laboratories and across platforms. **Nat Methods**, 2(351)–6. [PubMed: 15846362]
[7] BARAHONA, G. V. (2018): **Modelo estadístico pedagógico para la toma de decisiones administrativas y académicas con impacto en el mejoramiento continuo del rendimiento de los estudiantes universitarios, basado en los métodos de selección CUR** (Doctoral dissertation, Universidad de Salamanca).
[8] BARAHONA, G. V., BARREIRO, C. M. M., GARCÍA, N. G., GONZÁLEZ, S. H., BARBA, M. S. AND VILLARDÓN, M. P. G. (2019): DYNAMIC CUR, AN ALTERNATIVE TO VARIABLE SELECTION IN CUR DECOMPOSITION. **REVISTA INVESTIGACION OPERACIONAL**, 40(3), 391-399.
[9] BELHADJI, A., BARDENET, R. and CHAINAIS, P. (2020): A determinantal point process for column subset selection. **The Journal of Machine Learning Research**, 21(1), 8083-8144.
[10] 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.
[11] 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.
[12] BOLÓN-CANEDO, V., SETH, S., SÁNCHEZ-MAROÑO, N., ALONSO-BETANZOS, A. and PRINCIPE, J. (2011): Statistical dependence measure for feature selection in microarray datasets, in: **19th European Symposium on Artificial Neural Networks-ESANN**, 23–28.
[13] BODOR, A. and SOLYMOSI, N. (2012): rCUR: CUR decomposition package. **R package version 1.3**. https://CRAN.R-project.org/package=rCUR.
[14] BODOR, A., CSABAI, I., MAHONEY, M. W. and SOLYMOSI, N. (2012): rCUR: an R package for CUR matrix decomposition. **BMC Bioinformatics**, 13, 1-6.
[15] BOULESTEIX, A. L. and STRIMMER, K. (2007): Partial least squares: a versatile tool for the analysis of high-dimensional genomic data. **Briefings in bioinformatics**, 8(1), 32-44.
[16] BOUTSIDIS, C. (2011): Topics in matrix sampling algorithms. **arXiv preprint arXiv**:1105.0709.
[17] BOUTSIDIS, C. and WOODRUFF, D. P. (2014): Optimal CUR matrix decompositions. In **Proceedings of the forty-sixth annual ACM symposium on Theory of computing**, 353-362.
[18] BOUTSIDIS, C., MAHONEY, M. W. and DRINEAS, P. (2008a): On selecting exactly k columns from a matrix, **Submitted for publication**. https://www.researchgate.net/profile/Christos-Boutsidis/publication/228451455_On_selecting_exactly_k_columns_from_a_matrix/links/02e7e52aea2e80833a000000/On-selecting-exactly-k-columns-from-a-matrix.pdf.
[19] BOUTSIDIS, C., MAHONEY, M. W. and DRINEAS, P. (2008b): Unsupervised feature selection for principal components analysis. In **Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining,** 61-69.

[20] 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.
[21] Broad institute. Cancer Program Data Sets. Available in http://www.broadinstitute.org/cgi-bin/cancer/datasets.cgi. **Consulted** September, 2017.
[22] CARREIRA-PERPINÁN, M. A. (1997): A review of dimension reduction techniques. Department of Computer Science. University of Sheffield. **Tech. Rep. CS-96-09**, 9, 1-69.
[23] CARTER, S. L., EKLUND, A. C., MECHAM, B. H., KOHANE, I. S. AND SZALLASI, Z. (2005): Redefinition of Affymetrix probe sets by sequence overlap with cDNA microarray probes reduces cross-platform inconsistencies in cancer-associated gene expression measurements. **BMC Bioinformatics**, 6(107). [PubMed: 15850491].
[24] CHEN, D., QIAN, G., SHI, C. and PAN, Q. (2017): Breast cancer malignancy prediction using incremental combination of multiple recurrent neural networks. In **Proceedings of international conference on neural information processing**, 43–52.
[25] CLEOFAS-SÁNCHEZ, L., SÁNCHEZ, J. S. and GARCÍA, V. (2019): Gene selection and disease prediction from gene expression data using a two-stage hetero-associative memory. **Progress in Artificial Intelligence**, 8 (1), 63–71.
[26] DRAGHICI, S., KHATRI, P., EKLUND, A. C. and SZALLASI, Z. (2006): Reliability and reproducibility issues in DNA microarray measurements. **Trends Genet**, 22(101)–9. [PubMed: 16380191].
[27] DRINEAS, P. and MAHONEY, M. W. (2016): RandNLA: randomized numerical linear algebra. **Communications of the ACM**, 59(6), 80-90.
[28] DRINEAS, P., MAHONEY, M. W. and MUTHUKRISHNAN, S. (2006a): Polynomial time algorithm for column-row based relative-error low-rank matrix approximation. **DIMACS TR**: 2006-04, 1-15.
[29] DRINEAS, P., MAHONEY, M. W. and MUTHUKRISHNAN, S. (2006b): Subspace Sampling and Relative-Error Matrix Approximation: Column-Based Methods. In **Approximation, Randomization and Combinatorial Optimization. Algorithms and Techniques**, 316-326. Springer, Berlin, Heidelberg.
[30] DRINEAS, P., MAHONEY, M. W. and MUTHUKRISHNAN, S. (2008): Relative-error CUR matrix decompositions. **SIAM Journal on Matrix Analysis and Applications**, 30, 844-881.
[31] DOUGLAS, B. and MARTIN, M. (2017): Matrix: Sparse and Dense Matrix Classes and Methods. **R package version 1.2-12**. https://CRAN.R-project.org/package=Matrix.
[32] Feature Selection Datasets at Arizona State University. Available in http://featureselection.asu.edu/datasets.php **Consulted** September, 2017.
[33] FREIJE, W., CASTRO-VARGAS, F., FANG, Z., HORVATH, S., CLOUGHESY, T., LIAU, L., MISCHEL, P. and NELSON, S. (2004): Gene expression profiling of gliomas strongly predicts survival, **Cancer Res.**, 64(18), 6503–6510.
[34] GEMAN, O., CHIUCHISAN, I., COVASA, M., DOLOC, C., MILICI, M. R. and MILICI, L. D. (2016): Deep learning tools for human microbiome big data. In **Proceedings of international workshop soft computing applications**, 265–275.
[35] GUI, J., WANG, S. L. and LEI, Y. K. (2010): Multi-step dimensionality reduction and semi-supervised graph-based tumor classification using gene expression data. **Artificial intelligence in medicine**, 50(3), 181-191.
[36] HALL, M. (1999): **Correlation-Based Feature Selection for Machine Learning**. PhD thesis, Citeseer.
[37] HALL, M. AND SMITH, L. (1998): Practical feature subset selection for machine learning. **Comput. Sci**, 98, 181–191.
[38] 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.
[39] HIRA, Z. M. and GILLIES, D. F. (2015): A Review of Feature Selection and Feature Extraction Methods Applied on Microarray Data. **Advances in Bioinformatic**s, 2015, 24-36.
[40] HELLER, M. J. (2002): DNA microarray technology: devices, systems, and applications. **Annual review of biomedical engineering**, 4(1), 129-153.
[41] HORNIK, K., BUCHTA, C. and ZEILEIS, A. (2009): Open-Source Machine Learning: R Meets Weka. **Computational Statistics**, 24(2), 225–232. doi:10.1007/s00180-008-0119-7.
[42] INZA, I., LARRANAGA, P., BLANCO, R., and CERROLAZA, A. J. (2004): Filter versus wrapper gene selection approaches in DNA microarray domains. **Artificial intelligence in medicine**, 31(2), 91-103.
[43] JARVINEN, A. K., HAUTANIEMI, S., EDGREN, H., AUVINEN, P., SAARELA, J., KALLIONIEMI, O. P., ***et al.*** (2004): Are data from different gene expression microarray platforms comparable? **Genomics**, 83(1164)–8. [PubMed: 15177569].

[44] JIA, W., SUN, M., LIAN, J. and HOU, S. (2022): Feature dimensionality reduction: a review**. Complex & Intelligent Systems**, 8(3), 2663-2693.
[45] JENSSEN, T. K., LANGAAS, M., KUO, W. P., SMITH-SORENSEN, B., MYKLEBOST, O. and HOVIG, E. (2002): Analysis of repeatability in spotted cDNA microarrays. **Nucleic Acids Res**, 30(3235)–44. [PubMed: 12136105]
[46] 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.
[47] JOLLIFFE, I. T. (2002): **Principal component analysis for special types of data** (pp. 338-372). Springer New York.
[48] KARATZOGLOU, A., SMOLA, A. and HORNIK, K. (2023): kernlab: Kernel-Based Machine Learning Lab. **R package version 0.9-32**, https://CRAN.R-project.org/package=kernlab.
[49] Kent RidgeBio-Medical Dataset. Available in http://datam.i2r.a-star.edu.sg/datasets/krbd **Consulted** September, 2017.
[50] KISHORE KUMAR, N. and SCHNEIDER, J. (2017): Literature survey on low rank approximation of matrices. **Linear and Multilinear Algebra**, 65(11), 2212-2244.
[51] KONONENKO, I. (1994): Estimating attributes: analysis and extensions of relief, in: **Machine Learning: ECML-94**, Springer, 171–182.
[52] KUHN, M. (2008): Building Predictive Models in R Using the caret Package. **Journal of Statistical Software**, 28(5), 1–26. doi:10.18637/jss.v028.i05.
https://www.jstatsoft.org/index.php/jss/article/view/v028i05.
[53] KURUVILLA, F. G., PARK, P. J. and SCHREIBER, S. L. (2002): Vector algebra in the analysis of genomewide expression data**. Genome Biol** 3:research0011.1-0011.11.
[54] LAZAR, C., TAMINAU, J., MEGANCK, S., STEENHOFF, D., COLETTA, A., MOLTER, C., ... and NOWE, A. (2012): A survey on filter techniques for feature selection in gene expression microarray analysis. **IEEE/ACM transactions on computational biology and bioinformatics**, *9*(4), 1106-1119.
[55] 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
[56] LECUN, Y., BENGIO, Y. and HINTON, G. (2015): Deep learning. **Nature**, 521(7553), 436-444.
[57] LI, J., CHENG, K., WANG, S., MORSTATTER, F., TREVINO, R. P., TANG, J. and LIU, H. (2017): Feature selection: A data perspective. **ACM computing surveys (CSUR)**, 50(6), 1-45.
[58] MAHONEY, M. W. (2011). Randomized algorithms for matrices and data. **Foundations and Trends® in Machine Learning**, 3(2), 123-224.
[59] MAHONEY, M. W. and DRINEAS, P. (2009): CUR matrix decompositions for improved data analysis. **Proceedings of the National Academy of Sciences**, 106(3), 697-702.
[60] MEDSKER, L. R. and JAIN, L. (2001): Recurrent neural networks. **Design and Applications**, 5(64-67), 2.
[61] NIELSEN, T. *et al.* (2002): Molecular characterisation of soft tissue tumours: A gene expression study**. Lancet**, 359,1301–1307.
[62] 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.
[63] 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(8), 1226–1238.
[64] PETRICOIN, E., ARDEKANI, A., HITT, B., LEVINE, P., FUSARO, V., STEINBERG, S., MILLS, G., SIMONE, C., FISHMAN, D., KOHN, E. ***et al.*** (2002): Use of proteomic patterns in serum to identify ovarian cancer, **Lancet**, 359(9306), 572–577.
[65] POMEROY, S., TAMAYO, P., GAASENBEEK, M., STURLA, L., ANGELO, M., MCLAUGHLIN, M., KIM, J., GOUMNEROVA, L,. BLACK, P., LAU, C., ***et al*** (2002): Prediction of central nervous system embryonal tumour outcome based on gene expression, **Nature**, 415(6870), 436–442.
[66] R CORE TEAM (2017): R: A language and environment for statistical computing. **R Foundation for Statistical Computing**, Vienna, Austria. https://www.R-project.org/
[67] RAMÍREZ-SALCEDO, J., CHÁVEZ, L., SANTILLÁN-TORREZ, J. and GUZMÁN-LEÓN, S. (2014): Microarreglos de DNA: Fabricación, Proceso y Análisis. Herramientas Moleculares Aplicadas en Ecología: Aspectos Teóricos y Prácticos, **Secretaría de Medio Ambiente y Recursos Naturales: Tlalpan**, México, 203-229.
[68] REVOLUTION ANALYTICS and WESTON, S. (2015): foreach: Provides Foreach Looping Construct for R. **R package version 1.4.3**. https://CRAN.R-project.org/package=foreach.

[69] REYES-NAVA, A., CRUZ-REYES, H., ALEJO, R., RENDÓN-LARA, E., FLORES-FUENTES, A. A. and GRANDA-GUTIÉRREZ, E. E. (2019): Using deep learning to classify class imbalanced gene-expression. In **Proceedings of Iberoamerican congress on pattern recognition (CIARP): 1**, 46–54. 10.1007/978-3-030-13469-3.
[70] RODRÍGUEZ, Y. E. T. (2023): RESTRICTED CUR MATRIX DECOMPOSITION: A NOVEL TECHNIQUE FOR GENES SUBSET SELECTION IN PROBLEMS OF CLASSIFICATION OF CANCER TUMORS. **REVISTA INVESTIGACIÓN OPERACIONAL**, 44(2), 281-293.
[71] RODRÍGUEZ, Y. E. T., ONES, V. G., GARCÍA, J. E. S. and VELAR, R. C. (2012): Utilización combinada de métodos exploratorios y confirmatorios para el análisis de la actividad antibacteriana de la cefalosporina (Parte I). **REVISTA INVESTIGACIÓN OPERACIONAL**, 33(3), 47-55.
[72] SOUTHERN, E. M. (1975): Detection of specific sequences among DNA fragments separated by gel electrophoresis. **J Mol Biol**, 98(3), 503-517.
[73] SPIRA, A., BEANE, J., SHAH, V., STEILING, K., LIU, G., SCHEMBRI, F., GILMAN, S., DUMAS, Y., CALNER, P., SEBASTIANI, P. ***et al***. (2007): Airway epithelial gene expression in the diagnostic evaluation of smokers with suspect lung cancer, **Nat. Med.**, 13(3), 361–366.
[74] TARCA, A. L., ROMERO, R. and DRAGHICI, S. (2006): Analysis of microarray experiments of gene expression profiling. **Am J Obstet Gynecol.**; 195(2), 373–388.
[75] THE INTERNATIONAL HAPMAP CONSORTIUM. (2005): A haplotype map of the human genome. **Nature**, 437, 1299-1320.
[76] VENABLES, W. N. and RIPLEY, B. D. (2002): **Modern Applied Statistics with S**. Fourth Edition. Springer, New York. ISBN 0-387-95457-0.
[77] VILORIA, A., BONERGE, O., LEZAMA, P. and MERCADO-CARUZO, N. (2020): Unbalanced data processing using oversampling: Machine learning unbalanced data processing using oversampling: Machine learning. **Procedia Computer Science**, 175, 108–113. 10.1016/j.procs.2020.07.018.
[78] WANG, N. N. (2009): **Statistical Problems in DNA Microarray Data Analysis** (Doctoral dissertation, UC Berkeley).
[79] WEIHS, C., LIGGES, U., LUEBKE, K. and RAABE, N. (2005): klaR Analyzing German Business Cycles. In Baier D, Decker R, Schmidt-Thieme L (eds.), **Data Analysis and Decision Support**, 335-343.
[80] WICKHAM, H. (2009): **ggplot2: Elegant Graphics for Data Analysis**. Springer-Verlag New York.
[81] YANG, J., RUBEL, O., PRABHAT, MAHONEY, M. W. and BOWEN, B. P. (2015). Identifying important ions and positions in mass spectrometry imaging data using CUR matrix decompositions. **Analytical chemistry**, 87(9), 4658-4666.
[82] YU, L. and LIU, H. (2003): Feature selection for high-dimensional data: a fast correlation-based filter solution, in: **Machine Learning-International Workshop then Conference-**, 20, 856.
[83] ZHAO, Z. and LIU, H. (2007): Searching for interacting features, in: **Proceedings of the 20th International Joint Conference on Artifical Intelligence**, Morgan Kaufmann Publishers Inc., 1156–1161.