BioData Mining

METHODOLOGY Open Access

A new pipeline for structural characterization and classification of RNA-Seq microbiome data

Sebastian Racedo[1], Ivan Portnoy[1,2*], Jorge I. Vélez[1], Homero San-Juan-Vergara[1], Marco Sanjuan[1] and Eduardo Zurek[1]

* Correspondence: iportnoy@cuc.edu.co; iportnoy@uninorte.edu.co
[1]Universidad del Norte, Barranquilla, Colombia
[2]Productivity and Innovation Department, Universidad de la Costa, Calle 58 # 55-56, Barranquilla, Colombia

Abstract

Background: High-throughput sequencing enables the analysis of the composition of numerous biological systems, such as microbial communities. The identification of dependencies within these systems requires the analysis and assimilation of the underlying interaction patterns between all the variables that make up that system. However, this task poses a challenge when considering the compositional nature of the data coming from DNA-sequencing experiments because traditional interaction metrics (e.g., correlation) produce unreliable results when analyzing relative fractions instead of absolute abundances. The compositionality-associated challenges extend to the classification task, as it usually involves the characterization of the interactions between the principal descriptive variables of the datasets. The classification of new samples/patients into binary categories corresponding to dissimilar biological settings or phenotypes (e.g., control and cases) could help researchers in the development of treatments/drugs.

Results: Here, we develop and exemplify a new approach, applicable to compositional data, for the classification of new samples into two groups with different biological settings. We propose a new metric to characterize and quantify the overall correlation structure deviation between these groups and a technique for dimensionality reduction to facilitate graphical representation. We conduct simulation experiments with synthetic data to assess the proposed method's classification accuracy. Moreover, we illustrate the performance of the proposed approach using Operational Taxonomic Unit (OTU) count tables obtained through 16S rRNA gene sequencing data from two microbiota experiments. Also, compare our method's performance with that of two state-of-the-art methods.

Conclusions: Simulation experiments show that our method achieves a classification accuracy equal to or greater than 98% when using synthetic data. Finally, our method outperforms the other classification methods with real datasets from gene sequencing experiments.

Keywords: Microbial communities, Compositional nature, Classification method, 16 rRNA sequencing

Background

Microorganisms living inside and on humans are known as the microbiota. When integrated with their genes' information, it is known as the microbiome. The Human Microbiome Project (HMP) was an endeavor for the characterization of the human microbiota to further understanding its impact on human health and diseases [1].

In recent years, biological sciences have experienced substantial technological advances that have led to the rediscovery of systems biology [2–4]. These advances were possible thanks to the technological ability to completely sequence the genome from any organism at a low cost [5, 6]. Such advances triggered the development of various analytic approaches and technologies to simultaneously monitoring all the components within cells (e.g., genes and proteins). With the genome information and analytic technologies, the mining and exploration of the resulting data opened up the possibility to better understand biological systems, such as microbial populations, and their complexity. The network structure of such biological systems can give insight into the underlying interactions taking place within those systems [7–10]. Furthermore, the understanding of these interactions can lead to the discovery of new methods that can help physicians, biologists, scientists, and healthcare workers with disease diagnosis, gene identification, classification of new data, and many other tasks [11].

We initially conducted a literature search in different medical, biological, and engineering databases as well as academic sites prestigious journals such as BMC Bioinformatics, PLOS ONE, ScienceDirect, and IEEE Xplore using the queries "correlation structure for gene expression classifications," "classifiers for compositional data," and "classifiers based on correlation structures" in order to identify papers in English using procedures for sample classification based on correlation structures in the 2009–2019 time window. Figure 1 shows the evolution of the number of publications retrieved when the keywords "correlation structure for gene expression classifications" are used. Publications were retrieved from several academic sites, namely BMC Bioinformatics, PLOS One, ScienceDirect, and Scopus. Figure 2 summarizes the current principal stages of gene expression analysis for sample classification.

Operational Taxonomic Unit (OTU) count tables are the usual output when processing the 16S rRNA sequences of microbiota samples [12]. These tables show the relative abundances of the bacteria that make a microbiota population (e.g., the human gut microbiota). OTU-based data have a compositional nature, which makes them difficult to work with [13, 14]. Thus, data transformation is required prior to any further analysis.

Aitchison [15] proposed two transformations to compensate for the data's compositionality, thus allowing the use of standard metrics in further analysis. The first transformation is the additive log-ratio (alr), which is defined as:

$$alr(\boldsymbol{x}) = \left(\ln \frac{x_1}{x_j}, \ldots, \ \ln \frac{x_{j-1}}{x_j}, \ \ln \frac{x_{j+1}}{x_j}, \ \ln \frac{x_n}{x_j} \right) \tag{1}$$

where x_j is an element of $\{x_1, x_2, x_3 \ldots, x_n\}$. Because one value x_j is selected as the denominator to build the log-ratios, the alr has been criticized as being subjective since the outcome depends mostly on the value of x_j selected [15–18].

The second transformation proposed by Aitchison is the centered log-ratio (clr), which is defined as:

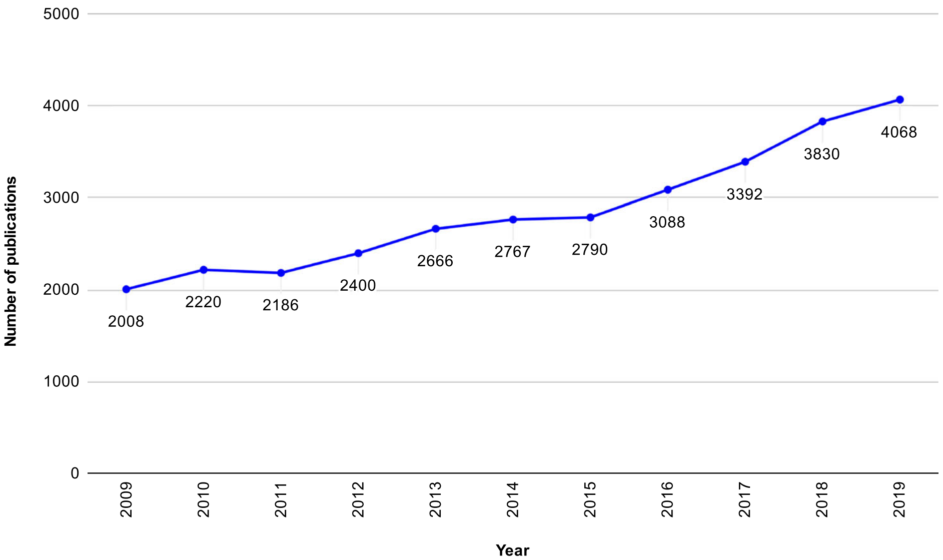

Fig. 1 Evolution of the number of publications per year from 2009 to 2019

$$clr(\boldsymbol{x}) = \left[\ln \frac{x_1}{g(\boldsymbol{x})}, \ln \frac{x_2}{g(\boldsymbol{x})}, \ldots, \ln \frac{x_n}{g(\boldsymbol{x})} \right] \tag{2}$$

where $g(\boldsymbol{x}) = \left(\prod_{i=1}^{n} x_i\right)^{\frac{1}{n}}$ is the geometric mean. The use of $g(\boldsymbol{x})$ avoids the subjectivity of the alr transformation since the method is taking all the information of $\boldsymbol{x}$ [15–19]. The clr transformation has proven to be reliable and has been extensively used in the scientific literature over the years to analyze microbiome data.

In [20] authors proposed a transformation called the isometric log-ratio (irl) transformation. This approach takes any compositional data $\boldsymbol{x} \in S^N$, and computes $ilr(\boldsymbol{x}) = z = [z_1, z_2, \ldots, z_N]$, where z_i is calculated as:

$$z_i = \sqrt{\frac{N-i}{N-i+1} \ln \left(\frac{x_i}{\sqrt[N-i]{\prod_{j=i+1}^{N} x_j}} \right)}, \quad i = 1, .., N. \tag{3}$$

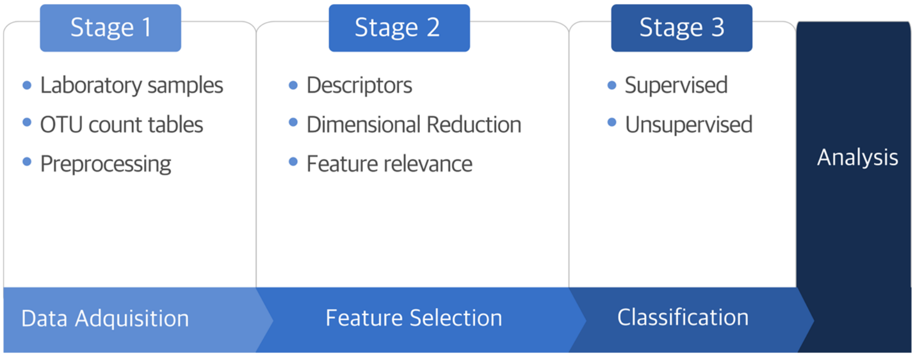

Fig. 2 Scheme of gene analysis used for sample classification

However, implementing the ilr transformation poses serious practical difficulties for high-dimension data as the computational complexity increases rapidly with dimensionality [21].

Feature selection

After transforming the data, the next step is to separate the data into train, test, and validation sets, although in some cases only the train and test sets are considered. One of the most common problems prior to that step is the limitation of the number of data samples. Indeed, for a normal classifier to be employed using multivariate metrical techniques, the sample size required for optimum training is in order of thousands. This is known as the “curse of dimensionality” problem, and the usual way to overcome this limitation is by using a dimensionality reduction technique to collapse all the attributes (variables) into a lower-dimension space where the most dominant information of the dataset can be retrieved [13, 22].

Feature selection methods are usually separated into three categories: filter, wrapper, and embedded. Table 1 summarizes different approaches for feature selection in gene expression data, the most relevant categories for feature selection, and the current weaknesses when analyzing gene expression data. Filter methods can work with univariate and multivariate data, where univariate methods focus on each feature separately and multivariate methods focus on finding relationships between features [23, 24]. Here we only consider multivariate methods.

The abovementioned filter methods tend to be computationally efficient. Wrapper methods, on the other hand, tend to have a better performance in selecting features since they take a model hypothesis into account, meaning that a training and testing procedure is made in the feature space. However, this approach is computationally inefficient and is more problematic as the feature space grows [23, 26, 29, 30]. Embedded methods make the feature selection based on the classifier (i.e., selected features might not work with any other classifier) and hence tend to have a better computational performance than wrappers. This is the case because the optimal set of descriptors is built when the classifier is constructed and the feature selection is affected by the hypotheses made by the classifier [23, 26, 29–31].

In [14], authors presented SParse InversE Covariance Estimation for Ecological ASSociation Inference (SPIEC-EASI), a novel strategy to infer networks from a high dimensional community compositional data. SPIEC-EASI estimates the interaction graph from the transformed data using either Recursive Feature selection or Sparse Inverse Covariance selection and seeks to infer an underlying graphical model using conditional independence. In [32] authors proposed a modification of the Support Vector Machine

Table 1 Summary of feature selection approaches in gene expression analysis

Category	Description	Weaknesses	References
Filter	- Extract features from the data without any type of learning involved.	- Ignore interaction with the classifier.	[13, 23, 25–30]
Wrapper	- Use learning approaches to evaluate which features are useful.	- Risk of overfitting. - Classifier dependent selection.	[23, 26, 29, 30]
Embedded	- Combine the traditional feature selection step with the classifier construction.	- Classifier dependent selection.	[23, 26, 29–31]

– Recursive Feature Elimination (SVM-RFE) algorithm for feature selection. SVM-RFE removes one irrelevant feature at each iteration, but this can be troublesome when the number of features is large. Thus, its modification, namely Correlation based Support Vector Machine – Recursive Multiple Feature Elimination (CSVM-RMFE), finds the correlated features and removes more than one irrelevant feature per iteration. Rao and S. Lakshminarayanan [13] presented a new significant attribute selection method based on the Partial Correlation Coefficient Matrix (PCCM).

Classification

The final step after finding the most relevant features of the transformed data is to select a classifier. In clinical and bioinformatic research, prediction models are extensively used to derive classification rules useful to accurately predict whether a patient has or would develop a disease, whether the treatment is going to work, or even whether a disease would recur [33–35]. Table 2 summarizes the relevant aspects of some widely used classifiers.

Depending on the data, a classifier can belong to one of two groups: supervised or unsupervised [36]. In supervised classification (learning), samples are labeled according to some a priori-defined classes or categories, whereas in unsupervised learning, samples are not labeled, and the classifier clusters the data into different classes or categories after maximizing or minimizing a set of criteria.

Dembélé and Kastner [37] presented a new Fold Change method that can detect differentially expressed genes in microarray data. The traditional fold change method works by calculating the ratio between the averages from the samples (usually two different biological conditions, e.g., control and case samples). Then, cutoff values (e.g., 0.5 for down- and 2 for up-regulated) are used to select genes under/above such thresholds. This new approach is more accurate and faster than the traditional method and can assign a metric to each differentially expressed gene, which can be used as a selection criterion.

Belciug and F. Gorunescu [43] proposed a novel initialization of a single hidden layer feedforward neural network's input weights using the knowledge embedded in the connections between variables and class labels. The authors expressed this by the non-parametric Goodman-Kruskal Gamma rank correlation instead of the traditional random initialization. The use of this correlation also helped to increase computational speed by eliminating unnecessary features based on the significance of the rank correlation between variables and class labels.

Table 2 Summary of classifiers used in gene expression analysis

Category	Classifier	References
Metrical and classical	- Probabilistic: Bayesian classifier, probabilistic linear discriminant analysis. - Non probabilistic: Support Vector Machine (SVM), SVM-RFE, Nearest-neighbor (NN), linear discriminant analysis.	[13, 37–41]
Artificial Intelligence	- Fuzzy Logic, Genetic Algorithms, Classification and Regression trees.	[13, 38, 39, 42, 43]
Boosting	- LogitBoost, AdaBoost.M1, GradientBoosting (GrBoost)	[13, 14, 38, 39, 44]

In [42], authors proposed a framework to find information about genes and to classify gene combinations belonging to its relevant subtype using fuzzy logic, which adapts numerical data (input/output pairs) into human linguistic terms, offering good capabilities to deal with noisy and missing data. However, defining the rules and membership functions might require a lot of prior knowledge from a human expert [41]. Dettling and P. Bühlmann [44] proposed a boosting method combining a dimensionality reduction step with the LogitBoost algorithm [45] and compared it to AdaBoost.M1 [46], the nearest neighbor classifier [47], and classification and regression trees (CART) using gene expression data [48]. Dettling and P. Bühlmann showed that, for low dimensional data, LogitBoost can perform slightly better than AdaBoost.M1, and that for real high dimensional data, their approach can outperform the other classifiers in some cases.

In this paper, we present a new method to classify samples into two groups with different characteristics (i.e., phenotypes, health condition, among others) when data of compositional nature is available. Our method relies on a new metric to quantitatively characterize the overall correlation structure deviation when comparing the two datasets and a new dimensionality reduction approach. The proposed method is assessed and compared, based on classification accuracy, to two state-of-the-art methods using both synthetic datasets and real datasets from RNA-16s sequencing experiments.

Proposed classification method

Here, we explain in detail the proposed classification method. First, in section "Data pretreatment", we introduce the Data Pretreatment stage, and in section "Assessing correlation structure distortion", a novel metric to be used as the metric to assess correlation structure distortion is described. Finally, in section "Dimensionality reduction technique", we present the proposed classification rule, which is based on the previously defined metric and a proposed dimensionality-reduction approach to assess the disruption of a dataset's correlation structure after a new sample is included.

Data pretreatment

Let $X_c^\rho \in \mathbb{R}^{n_c \times m}$ and $X_v^\rho \in \mathbb{R}^{n_v \times m}$ be the OTU count tables where m features are assessed in n_c and n_v samples from control and case individuals, respectively. In the expressions above, the superindex ρ indicates the datasets are 'raw' or without pretreatment. From now on, X_g^ρ will represent any of the two groups ($g = c$ for control, or $g = v$ for case).

When analyzing OTU counts tables, a log-ratio transformation, such as the clr, is to be applied [15, 18, 19] before estimating correlations. However, in order to apply the log-ratio transformation, it is necessary to consider that compositional count datasets may contain null values resulting from insufficiently large or non-existing samples. As log-ratio transformations require data with exclusively positive values, the use of a zero-replacement method is a must. Here we use the Bayesian-multiplicative (BM) algorithm proposed by Martín-Fernández [49]. Let $\boldsymbol{x}_{p_i} \in \mathbb{R}^{1 \times m}$ be the i-th row of the matrix X_g^ρ $(i = 1, 2, \ldots, n_g)$. The BM algorithm replaces the null counts by

$$BM\left(x_{p_{i,j}}\right) = \begin{cases} t_{i,j}\left(\dfrac{s_i}{n+s_i}\right), & \text{if } x_{p_{i,j}} = 0 \\ x_{p_{i,j}}\left(1 - \displaystyle\sum_{\forall k \ |x_{p_{i,j}}=0} t_{i,k}\left(\dfrac{s_i}{n+s_i}\right)\right), & \text{if } x_{p_{i,j}} \neq 0 \end{cases} \tag{4}$$

When using the Bayes-Laplace prior, we set $n = \sum_{j=1}^{m} x_{p_{i,j}}$, $t_{i,\ j} = m^{-1}$ and $s_i = m$. Let $X_g^{BM} \coloneqq BM(X_g^{\rho})$ be the resulting matrix after the BM algorithm is applied row-wise to X_g^{ρ}

.

To ensure the data's compositionality on X_g^{BM}, a closure operation [15, 18, 19] is applied to every row of X_g^{BM}, as follows:

$$c\left(\boldsymbol{x}_{p_i}^{BM}\right) = \frac{k}{\sum_{j=1}^{m} x_{p_{i,j}}^{BM}} \boldsymbol{x}_{p_i}^{BM} \tag{5}$$

where k is an arbitrary constant (usually $k = 100$). Let $X_g^{BM,c} \coloneqq c(BM(X_g^{\rho}))$ be the resulting matrix after the BM algorithm and the closure operation have been applied. Now, the clr transformation is applied to each vector $\boldsymbol{x}_p \in \mathbb{R}^{1 \times n}$ $X_g^{BM,c}$, as

$$clr\left(\boldsymbol{x}_p\right) = \left[\ \ln \frac{x_1}{g\left(\boldsymbol{x}_p\right)},\ \ln \frac{x_2}{g\left(\boldsymbol{x}_p\right)}, \ldots,\ \ln \frac{x_n}{g\left(\boldsymbol{x}_p\right)}\right] \tag{6}$$

where $g(x_p) = \left(\prod_{i=1}^{n} x_i\right)^{\frac{1}{n}}$ is the geometric mean. Hence,

$$X_g = clr\left(c\left(BM\left(X_g^{\rho}\right)\right)\right) \tag{7}$$

Finally, a normalization is applied, resulting in:

$$X_{g_{norm}} = \left(X_g - I_{n_g} b_g^T\right) \Sigma_g^{-1} \tag{8}$$

where $I_g = [1\ 1....1] \in \mathbb{R}^{n_g \times 1}$ is a column vector of ones, $b_g \in \mathbb{R}^{n_g \times 1}$ is a column vector that contains the means of all the variables in X_g, and $\Sigma_g \in \mathbb{R}^{m \times m}$ is a diagonal matrix that contains the standard deviation (σ_{g_i}, for $i = 1, \ldots,$ m) of all variables.

Assessing correlation structure distortion

Here, we introduce ϕ, a new metric to quantitatively assess the distortion in the correlation structure of a dataset after the incorporation of a new sample. The Pearson correlation matrix for X_g is calculated as follows [50]:

$$S_g = \frac{1}{n_g - 1} X_{g_{norm}}^T X_{g_{norm}} \tag{9}$$

Now, consider a new sample, $\boldsymbol{x}_p \in \mathbb{R}^{1 \times m}$. The pretreatment step for this sample yields:

$$\boldsymbol{x}_p = clr\left(c\left(BM\left(\boldsymbol{x}_p\right)\right)\right) \tag{10}$$

Let $\tilde{X}_g\mathbb{R}^{n_g\times m}$ be the (augmented) dataset X_g after incorporating the new sample, and let S_g and $\tilde{S}_g$ be the correlation matrices for X_g and $\tilde{X}_g$, respectively. The spectral decomposition for these matrices is

$$S_g = V_g\Lambda_g V_g^T, \qquad \tilde{S}_g = \tilde{V}_g\tilde{\Lambda}_g\tilde{V}_g^T \tag{11}$$

where

$$\Lambda_g = \begin{bmatrix} \lambda_{g_1} & & \\ & \ddots & \\ & & \lambda_{g_m} \end{bmatrix} \in\mathbb{R}^{m\times m},\ \ \tilde{\Lambda}_g = \begin{bmatrix} \tilde{\lambda}_{g_1} & & \\ & \ddots & \\ & & \tilde{\lambda}_{g_m} \end{bmatrix} \in\mathbb{R}^{m\times m} \tag{12}$$

are diagonal matrices containing the eigenvalues for S_g and $\tilde{S}_g$. Let $V_g = \begin{bmatrix} \boldsymbol{v}_{g_1} & \boldsymbol{v}_{g_2} & \cdots & \boldsymbol{v}_{g_m} \end{bmatrix} \in\mathbb{R}^{m\times m}$ and $\tilde{V}_g = \begin{bmatrix} \tilde{\boldsymbol{v}}_{g_1} & \tilde{\boldsymbol{v}}_{g_2} & \cdots & \tilde{\boldsymbol{v}}_{g_m} \end{bmatrix} \in\mathbb{R}^{m\times m}$ be the eigenvector matrices of S_g and $\tilde{S}_g$. Figure 3a illustrates, in a 2-dimensional example, the datasets X_g and $\tilde{X}_g$. Figure 3b illustrates the datasets after carrying out the pre-treatment, along with their eigenvectors (which are unitary) scaled by their corresponding eigenvalues obtained from the spectral decompositions. Note that scaled eigenvectors mark out the directions of largest variability, capturing high order interactions between the OTUs ruling the overall association structure. Therefore, looking at deviations in both the magnitude and direction of those scaled eigenvectors must give insightful information on overall changes in the association structure of a microbiota population.

Based on the abovementioned remarks, we introduce ϕ to characterize the distortion produced in the underlying correlation structure when two OTU counts datasets are compared. This metric first requires a dimensional reduction, which will be performed by selecting the principal components for each sample group. This procedure, integrated within the Principal Component Analysis (PCA) algorithm [25], consists of finding the minimum number of eigenvalues a_g or $\tilde{a}_g$ (for X_g and $\tilde{X}_g$, respectively) that explain 100(1 – α)% of the total variance, i.e.:

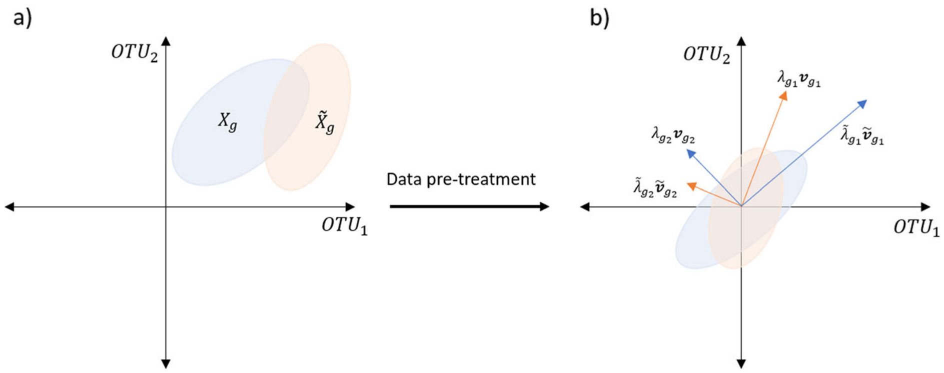
a)
OTU_2
X_g
$\tilde{X}_g$
OTU_1
Data pre-treatment
b)
OTU_2
$\lambda_{g_1}\boldsymbol{v}_{g_1}$
$\lambda_{g_2}\boldsymbol{v}_{g_2}$
$\tilde{\lambda}_{g_1}\tilde{\boldsymbol{v}}_{g_1}$
$\tilde{\lambda}_{g_2}\tilde{\boldsymbol{v}}_{g_2}$
OTU_1

Fig. 3 Bidimensional representation of datasets $\tilde{X}_g$ and X_g **a** without pretreatment, and **b** after the pretreatment along with the eigenvectors scaled by the corresponding eigenvalues

$$\frac{\sum_{i=1}^{a_g}\lambda_{g_i}}{\sum_{i=1}^{m}\lambda_{g_i}} \leq (1-\alpha), \qquad \frac{\sum_{i=1}^{\tilde{a}_g}\lambda_{g_i}}{\sum_{i=1}^{m}\lambda_{g_i}} \leq (1-\alpha) \tag{13}$$

Thus, ϕ is defined as

$$\phi = \sum_{j=1}^{\max\left(a_g,\tilde{a}_g\right)} \left[\max\left\{\lambda_{g_j},\tilde{\lambda}_{g_j}\right\} \left(\lambda_{g_j}-\tilde{\lambda}_{g_j}\right) \cos^{-1}\left(\boldsymbol{v}_{g_j}^T\tilde{\boldsymbol{v}}_{g_j}\right)\right] \tag{14}$$

where $(\lambda_{g_j}-\tilde{\lambda}_{g_j})$ is the algebraic difference (magnitude deviation) of the *j*-th eigenvalues in Λ_g and $\tilde{\Lambda}_g$, $\cos^{-1}(\boldsymbol{v}_{g_j}^T\tilde{\boldsymbol{v}}_{g_j})$ computes angular deviation between the *j*-th eigenvectors in V_g and $\tilde{V}_g$, and $\max\{\lambda_{g_j},\tilde{\lambda}_{g_j}\}$ provides a weighting factor so that the contribution of the *j*-th deviation to the index ϕ is proportional to the relative importance among principal components.

Dimensionality reduction technique

Now that we have a metric to measure the distortion caused in the correlation structure of the *g* group after the incorporation of a new sample, we could then infer to which group the new sample would belong, providing a classification criterion based on how distorted the correlation structure is when incorporating $\boldsymbol{x}_p$. The intuitive way of approaching the evaluation of the distortion would be to integrate $\boldsymbol{x}_p$ into X_g and (re)calculate the correlation matrix for the further evaluation of its distortion. However, considering that the *g* group may contain many samples, a single new sample may not be enough to generate a significant distortion in the correlation structure. Furthermore, if the number of samples in the groups is unbalanced, the distortion caused by the inclusion of a new sample may not be comparable.

An approach to overcome this dimensional problem is to randomly subsample a small number of rows in X_g, combining them with $\boldsymbol{x}_p$, and then calculating the distortion caused. This approach, however, would not include a considerable amount of information, which is contained in the rows that were left out. To address this issue, we propose a new dimensionality reduction approach that allows a weighted assessment of the distortion in S_g caused by the integration of a new sample $\boldsymbol{x}_p$. This approach will use all the information contained in the original data, with the objective of providing a classification algorithm for any upcoming sample.

The first step of the proposed approach is to find an expression for the distorted correlation matrix that reveals the natural weights of the contributions of X_g and $\boldsymbol{x}_p$ to the make-up of the new correlation structure. Suppose that the data is concatenated as:

$$\tilde{X}_g = \begin{bmatrix} X_g \\ \boldsymbol{x}_p \end{bmatrix} \mathbb{R}^{\tilde{n}_g\times m} \tag{15}$$

where $\tilde{n}_g = n_g + 1$ is the number of rows of $\tilde{X}_g$. Combining Eqs. (15) and (8) yields

$$\tilde{X}_g = \begin{bmatrix} X_{g_{norm}}\Sigma_g + I_{n_g} b_g^T \\ \boldsymbol{x}_p \end{bmatrix} \tag{16}$$

Normalizing $\tilde{X}_g$ produces

$$\tilde{X}_{g_{norm}} = \left(\tilde{X}_g - I_{\tilde{n}_g}\tilde{b}_g^T\right)\tilde{\Sigma}_g^{-1} = \begin{bmatrix} \left(X_{g_{norm}}\Sigma_g - I_{n_g}\Delta b_g^T\right)\tilde{\Sigma}_g^{-1} \\ \boldsymbol{x}_{p_{norm}} \end{bmatrix} \tag{17}$$

where $\tilde{b}_g$ is the vector that contains the means of $\tilde{X}_g$, $\tilde{\Sigma}_g$ is a diagonal matrix that contains the distorted standard deviations, $\Delta b_g \coloneqq \tilde{b}_g - b_g$ is the distortion in the mean vector, and $\boldsymbol{x}_{p_{norm}} = (\boldsymbol{x}_p - \tilde{b}_g^T)\tilde{\Sigma}_g^{-1}$. Both $\tilde{b}_g$ and $\tilde{\Sigma}_g$ are unknown. Thus, we need to derive expressions for them. The distorted means vector is calculated as $\tilde{b}_g = \frac{1}{\tilde{n}_g}\tilde{X}_g{}^T I_{\tilde{n}_g}$, which can be converted into:

$$\tilde{b}_g = \frac{n_g}{n_g+1}b_g + \frac{1}{n_g+1}\boldsymbol{x}_p^T \tag{18}$$

Equation (18) shows that the natural weights are $w_1 = \frac{n_g}{n_g+1}$ and $w_2 = \frac{1}{n_g+1}$ for b_g and $\boldsymbol{x}_p$, respectively. To find an expression for the diagonal matrix of distorted standard deviations, $\tilde{\Sigma}_g$, a column-wise subtraction of the mean vector for $\tilde{X}_g$ is performed:

$$\tilde{X}_{g_{mean-centered}} = \tilde{X}_g - I_{\tilde{n}_g}\tilde{b}_g^T = \begin{bmatrix} X_g - I_{n_g}\tilde{b}_g^T \\ \boldsymbol{x}_p - \tilde{b}_g^T \end{bmatrix} \tag{19}$$

Adding and subtracting $I_{n_g}b_g^T$ to $X_g - I_{n_g}\tilde{b}_g^T$ in Eq. (19) yields:

$$\tilde{X}_{g_{mean-centered}} = \begin{bmatrix} \left(X_g - I_{n_g}b_g^T\right) - I_{n_g}\Delta b_g^T \\ \boldsymbol{x}_p - \tilde{b}_g^T \end{bmatrix} \tag{20}$$

where

$$\tilde{X}_{g_{mean-centered}}(:,i) = \begin{bmatrix} \left(X_g(:,i) - b_g(i)I_{n_g}\right) - \Delta b_g(i)I_{n_g} \\ \boldsymbol{x}_p(i) - \tilde{b}_g(i) \end{bmatrix} \tag{21}$$

is the i-th column of $\tilde{X}_{g_{mean-centered}}(:,i)$, the corresponding i-th variable. Then, the variance of this i-th variable will be $\tilde{\sigma}_{g_i}^2 = \frac{1}{\tilde{n}_g-1}\left(\tilde{X}_{g_{mean-centered}}(:,i)\right)^T \tilde{X}_{g_{mean-centered}}(:,i)$, which can be written as:

$$\begin{aligned}(\tilde{n}_g-1)\tilde{\sigma}_{g_i}^2 = &\begin{bmatrix} \left(X_g^T(:,i) - b_g(i)I_{n_g}^T\right) - \Delta b_g(i)I_{n_g}^T & \boldsymbol{x}_p(i) - \tilde{b}_g(i) \end{bmatrix} \\ &\times \begin{bmatrix} \left(X_g(:,i) - b_g(i)I_{n_g}\right) - \Delta b_g(i)I_{n_g} \\ \boldsymbol{x}_p(i) - \tilde{b}_g(i) \end{bmatrix}\end{aligned} \tag{22}$$

Equation (22) can be further expanded as:

$$\begin{aligned}(\tilde{n}_g-1)\tilde{\sigma}_{g_i}^2 = &\left(X_g^T(:,i) - b_g(i)I_{n_g}^T\right)\left(X_g(:,i) - b_g(i)I_{n_g}\right) - \left(X_g^T(:,i) - b_g(i)I_{n_g}^T\right)\Delta b_g(i)I_{n_g} \\ &- \Delta b_g(i)I_{n_g}^T\left(X_g(:,i) - b_g(i)I_{n_g}\right) + \Delta b_c^2(i)I_{n_g}^T I_{n_g} + \left(\boldsymbol{x}_p(i) - \tilde{b}_g(i)\right)^2\end{aligned} \tag{23}$$

Notice that, in this expression, the terms $(X_g^T(:,i) - b_g(i)I_{n_g}^T)(X_g(:,i) - b_g(i)I_{n_g}) = (n_g - 1)\sigma_{g_i}^2$, $I_{n_g}^T I_{n_g} = n_g$, and $(X_g^T(:,i) - b_g(i)I_{n_g}^T)\Delta b_g(i)I_{n_g} = \Delta b_g(i)I_{n_g}^T(X_g(:,i) - b_g(i)I_{n_g})$. Then, Eq. (23) can be reduced to:

$$\begin{aligned}\left(\tilde{n}_g-1\right)\tilde{\sigma}_{g_i}^2 = {} & \left(n_g-1\right)\sigma_{g_i}^2-2\Delta b_g(i)I_{n_g}^T\left(X_g(:,i)-b_g(i)I_{n_g}\right)+n_g\Delta b_g^2(i) \\ & +\left(\boldsymbol{x}_p(i)-\tilde{b}_g(i)\right)^2\end{aligned} \tag{24}$$

Considering that $\tilde{n}_g = n_g+1$ and $I_{n_g}^T X_g(:,i) = I_{n_g}^T\left(b_g(i)I_{n_g}\right) = n_g b_g(i)$, it follows that

$$\tilde{\sigma}_{g_i} = \sqrt{\frac{n_g-1}{n_g}\sigma_{g_i}^2+\Delta b_g^2(i)+\frac{1}{n_g}\left(\boldsymbol{x}_p(i)-\tilde{b}_g(i)\right)^2} \tag{25}$$

From Eq. (25), notice that the (distorted) variances of the variables of the group $\tilde{X}_g$ depend on: (1) the original variances in X_g, with natural weight $\frac{n_g-1}{n_g}$; (2) the quadratic (mean centered) values of the new sample, $\left(\boldsymbol{x}_p(i)-\tilde{b}_g(i)\right)^2$, with natural weight $\frac{1}{n_g}$; and the quadratic values of the distortion in the mean vector, $\Delta b_g^2(i)$. Based on equation [25], the standard deviation matrix for all m variables is

$$\tilde{\Sigma}_g = \begin{bmatrix}\tilde{\sigma}_{g_1} & & \\ & \ddots & \\ & & \tilde{\sigma}_{g_m}\end{bmatrix} \tag{26}$$

Having expressions for $\tilde{b}_g$ and $\tilde{\Sigma}_g$, it follows that the distorted correlation matrix is calculated as $\tilde{S}_g = \frac{1}{\tilde{n}_g-1}\tilde{X}_{g_{norm}}^T\tilde{X}_{g_{norm}}$. Combining $\tilde{S}_g$ with Eq. (17) yields

$$\left(\tilde{n}_g-1\right)\tilde{S}_g = \begin{bmatrix}\tilde{\Sigma}_g^{-1}\left(\Sigma_g X_{g_{norm}}^T-\Delta b_g I_{n_g}^T\right) & \boldsymbol{x}_{p_{norm}}^T\end{bmatrix}\begin{bmatrix}\left(X_{g_{norm}}\Sigma_g-I_{n_g}\Delta b_g^T\right)\tilde{\Sigma}_g^{-1} \\ \boldsymbol{x}_{p_{norm}}\end{bmatrix} \tag{27}$$

It follows that,

$$\begin{aligned}\left(\tilde{n}_g-1\right)\tilde{S}_g = {} & \tilde{\Sigma}_g^{-1}\Sigma_g X_{g_{norm}}^T X_{g_{norm}}\Sigma_g\tilde{\Sigma}_g^{-1}-\tilde{\Sigma}_g^{-1}\Sigma_g X_{g_{norm}}^T I_{n_g}\Delta b_g^T\tilde{\Sigma}_g^{-1}-\tilde{\Sigma}_g^{-1}\Delta b_g I_{n_g}^T X_{g_{norm}}\Sigma_g\tilde{\Sigma}_g^{-1} \\ & +\tilde{\Sigma}_g^{-1}\Delta b_g I_{n_g}^T I_{n_g}\Delta b_g^T\tilde{\Sigma}_g^{-1}+\boldsymbol{x}_{p_{norm}}^T\boldsymbol{x}_{p_{norm}}\end{aligned} \tag{28}$$

As $X_{g_{norm}}^T X_{g_{norm}} = \left(n_g-1\right)S_g$, $\Sigma_g X_{g_{norm}}^T = X_g^T-b_g I_{n_g}^T$, $X_{g_{norm}}\Sigma_g = X_g-I_{n_g}b_g^T$, this expression can be expressed as:

$$\begin{aligned}\left(\tilde{n}_g-1\right)\tilde{S}_g = {} & \left(n_g-1\right)\tilde{\Sigma}_g^{-1}\Sigma_g S_g\Sigma_g\tilde{\Sigma}_g^{-1}-\tilde{\Sigma}_g^{-1}\left(X_g^T-b_g I_{n_g}^T\right)I_{n_g}\Delta b_g^T\tilde{\Sigma}_g^{-1}-\tilde{\Sigma}_g^{-1}\Delta b_g I_{n_g}^T\left(X_g-I_{n_g}b_g^T\right)\tilde{\Sigma}_g^{-1} \\ & +n_g\tilde{\Sigma}_g^{-1}\Delta b_g\Delta b_g^T\tilde{\Sigma}_g^{-1}+\boldsymbol{x}_{p_{norm}}^T\boldsymbol{x}_{p_{norm}}\end{aligned} \tag{29}$$

Now, as $X_g^T I_{n_g} = b_g I_{n_g}^T I_{n_g} = I_{n_g}^T X_g = I_{n_g}^T I_{n_g} b_g^T = n_g b_g$, the second and third terms of Eq. (29) disappear. Then, the distorted correlation matrix $\tilde{S}_g$ is given by

$$\tilde{S}_g = \frac{n_g-1}{n_g}\tilde{\Sigma}_g^{-1}\Sigma_g S_g\Sigma_g\tilde{\Sigma}_g^{-1}+\tilde{\Sigma}_g^{-1}\Delta b_g\Delta b_g^T\tilde{\Sigma}_g^{-1}+\frac{1}{n_g}\boldsymbol{x}_{p_{norm}}^T\boldsymbol{x}_{p_{norm}} \tag{30}$$

Note that, in this expression, $\tilde{S}_g$ depends on three terms:

1. $\tilde{\Sigma}_g^{-1}\Sigma_g S_g\Sigma_g\tilde{\Sigma}_g^{-1}$, which considers the contributions made from the non-distorted correlation matrix S_g after an actualization of the standard deviation, with a natural weight of $\frac{n_g-1}{n_g}$.

2. $\boldsymbol{x}_{p_{norm}}^T \boldsymbol{x}_{p_{norm}}$, which considers the contribution of the new sample to the constitution of the distorted correlation matrix, with a natural weight of $\frac{1}{n_g}$.
3. $\tilde{\Sigma}_g^{-1} \Delta b_g \Delta b_g^T \tilde{\Sigma}_g^{-1}$, which considers the effects of the distortion of Σ_g and b_g in $\tilde{S}_g$.

Finally, the distortion of the correlation matrix will be measured with the estimation of the deviation between S_g and $\tilde{S}_g$, using the metric $\phi(S_g, \tilde{S}_g)$ defined in Eq. (14). As previously mentioned, if the number of samples for the group g is large, the integration of x_p will barely cause a distortion in the correlation structure, even if it has different features compared to the samples in X_g. For example, if X_g were composed of 200 samples, the natural relative weight of the mean vector (b_c) for the construction of the distorted mean vector would be ~ 0.995, while the natural weight of the sample would (only) be ~ 0.005.

On the other hand, if the weights were calculated assuming that X_g is composed of few samples, that is, replacing n_g for n_g^{red} (so that $n_g^{red} < n_g$) in the quotients to calculate the relative weights, these weights would be more even and provide a weighting factor for the calculation of the distorted correlation matrix using all the information contained in the original samples of X_g (in b_g, Σ_g, and S_g). This is equivalent to finding a generatrix base of a few samples/patients (n_g^{red}) that can represent all the characteristics of X_g, incorporate $\boldsymbol{x}_p$, and then evaluate the distortion caused to the correlation structure, providing an artificial dimensional reduction. For example, if the relative weights were calculated assuming that X_g is composed only of three samples that exhibit all the attributes of the original dataset (i.e., $n_g^{red} = 3$), these weights would have the values of 0.75 and 0.25, respectively, for the calculation of the distorted mean vector.

The lower threshold for this artificial dimensional reduction could be found making $n_g^{red} = 2$ in the calculation of the relative weights. If $n_g^{red} = 1$, this would lead to leaving out all the information contained in S_g to the estimation of $\tilde{S}_g$ (see Eq. (30)). A similar result is obtained for the standard deviation (see Eq. (28)).

Proposed classification rule

Now that the artificial dimensional reduction approach has been proposed, it will be used alongside the metric ϕ for the creation of a tool to classify new samples/patients into either the control or case group. The classifier will work under the assumption that a sample's likelihood of belonging to either group is inversely proportional to the distortion caused by its incorporation into that group. This classification approach includes the following steps:

1. Store the new sample in $\boldsymbol{x}_p$.
2. Define the "maximum artificial dimension" to be evaluated as $n \leq \ min(n_c, n_v)$ $(n \in \mathbb{Z}^+)$. Choose a dimension "step of change", $\Delta n \in \mathbb{Z}^+$, such as $n - 2$ is divisible by Δn. Thus, $\frac{(n-2)}{\Delta n} + 1$ would define the number of artificial dimensions to be evaluated. Therefore, we set $n_g^{red} = (2, 2 + \Delta n, 2 + 2\Delta n, \ldots, n)$ for both $g = c$ and $g = v$.

3. Evaluate Eqs. (18), (25), (26) and (30) using n_g^{red} instead of n_g. Perform this evaluation for both $g = c$ and $g = v$, and for all values of n_g^{red}. Store the resulting distorted correlation matrices as

$$\tilde{S}_c = \left\{ \begin{matrix} \tilde{S}_{c|_{n_g^{red}=2}} \\ \vdots \\ \tilde{S}_{c|_{n_g^{red}=n}} \end{matrix} \right\}, \quad \tilde{S}_v = \left\{ \begin{matrix} \tilde{S}_{v|_{n_g^{red}=2}} \\ \vdots \\ \tilde{S}_{v|_{n_g^{red}=n}} \end{matrix} \right\} \tag{31}$$

4. For each $n_g^{red} = (2, 2 + \Delta n, 2 + 2\Delta n, \ldots, n)$, calculate

$$\left(\psi_g\right)\Big|_{n_g^{red}} \coloneqq \frac{1}{\left|\phi\left(S_g, \tilde{S}_{g_{n_g^{red}}}\right)\right|}, \qquad g = \{c, v\} \tag{32}$$

   where $|l|$ is the absolute value of l. In consequence, large values of ψ indicate a small distortion in the correlation structure, and therefore, a high degree of affinity between X_g and $\boldsymbol{x}_p$. On the other hand, small values of ψ indicate a big distortion and a low degree of affinity between X_g and $\boldsymbol{x}_p$.

5. Calculate the average value for $(\psi_g)|_{n_g^{red}}$ as

$$\overline{\psi}_g = \frac{1}{n} \sum_{\forall n_g^{red}} \left[\left(\psi_g\right)\Big|_{n_g^{red}} \right], \qquad g = \{c, v\} \tag{33}$$

6. Finally, the outcomes of the proposed classification rule, for a single sample, are $\overline{\psi}_c$ and $\overline{\psi}_v$. The method will classify the sample into the group with the greater value of $\overline{\psi}_g$. Figure 4 shows a graphical representation to visualize the outcome of the proposed classification method after classifying a set of new samples one-by-one.

Performance assessment with synthetic data

In this section, we assess the performance of the proposed method to correctly classify synthetically generated data.

Synthetic data generation

We conducted *in silico* experiments to assess the performance of the proposed method under different parameter settings. The following procedure was used to generate synthetic datasets:

1. Define the quadruplet $(n_i, m_j, \rho_c, \rho_v)$. Set $n = \{20,40,60,80,100,120,140,160\}$, $m = \{20,40,60,80,100,120,140\}$, $\rho_c = 0.1$, $\rho_v = 0.2$.

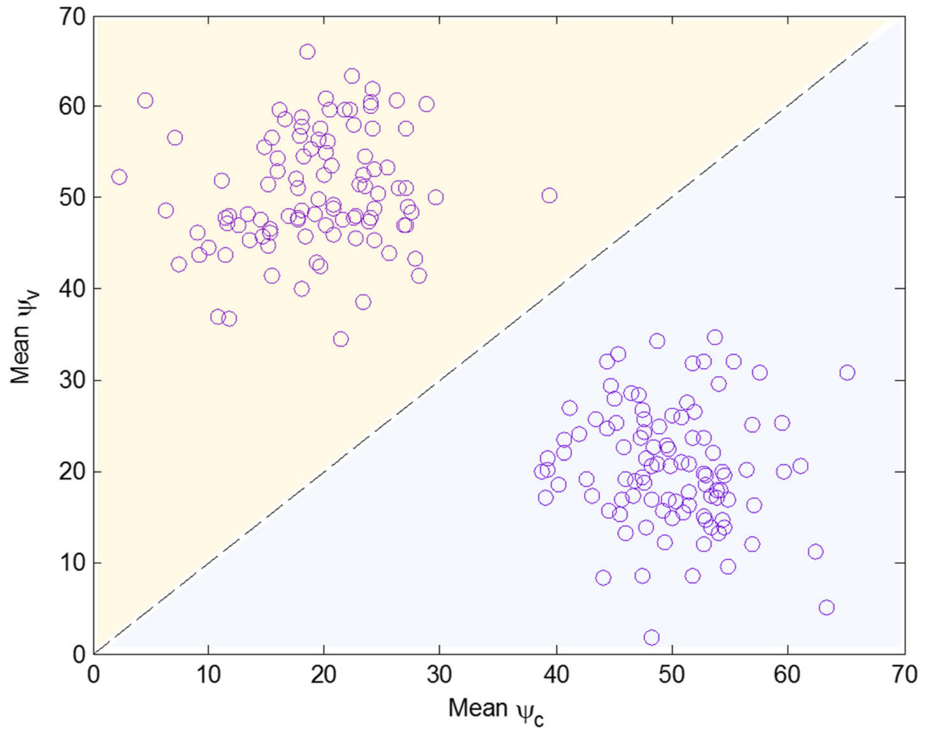

Fig. 4 Illustration of new samples and the line that separates both groups with the proposed method. Samples lying in the upper semi-plane will be classified in the case (*v*) group and in the control (*c*) group otherwise

2. For every quadruplet in step 1 construct a pair of generatrix correlation matrices, $\Sigma_{c_{j,c}}$ and $\Sigma_{v_{j,v}}$ as $\Sigma_{c_{j,c}} = (1-\rho_c)I_{m_j} + \rho_c 1_{m_j} 1_{m_j}^T$ and $\Sigma_{v_{j,v}} = (1-\rho_v)I_{m_j} + \rho_v 1_{m_j} 1_{m_j}^T$, where $I_{m_j}\in\mathbb{R}^{m_j\times m_j}$ is the identity matrix and $1_{m_j}\in\mathbb{R}^{m_j\times 1}$ is column vector of ones.
3. For every pair $(\Sigma_{c_{j,c}}, \Sigma_{v_{j,v}})$, B pairs of Normal-distributed matrices X_{c_r} and X_{v_r} (with $r = \{1, 2, \ldots, B\}$) of dimension $n_i \times m_j$ are generated. For this purpose, the NumPy [54] Python package was used. The number of experimental replicates was $B = 100$.

Performance assessment procedure

We used the correct classification rate (accuracy) as the assessment criterion to measure the performance of our method as follows:

1. Merge each (X_{c_r}, X_{v_r}) into a single matrix $X_{Total} = \begin{bmatrix} X_{c_r} \\ X_{v_r} \end{bmatrix} \in \mathbb{R}^{2n\times m}$.
2. For every pair (X_{c_r}, X_{v_r}), execute the proposed algorithm with each row sample $x_{p_i} = X_{Total_i}[i, :]$, $i = \{1, 2, \ldots, 2n\}$, and classify x_{p_i}.
3. Compute the average classification accuracy as:

$$\text{Accuracy} = 100 \times \frac{N}{2n} \tag{34}$$

where N is the number of correctly classified samples.

Performance assessment results with synthetic data

Table 3 summarizes the main results. Our method exhibits exceptional accuracy for all the configurations tested. Interestingly, accuracy decreases as the number of features m decreases and the sample size n increases.

Validation with real datasets

In this section, we study the performance of the proposed method using two real-world datasets, which contain OTU count tables obtained through 16S rRNA gene sequencing data from microbiota experiments. We also compare the classification accuracy of our method with those of two state-of-the-art methods: SVM [39] and SVM-RFE [41].

Datasets

The first dataset is from the American Gut Project (AGP) [51], which is one of the largest crowd-funded microbiome research projects. The second dataset is the Greengenes (GG) database [52], created with the PhyloChip 16s rRNA microarray. For the comparison experiment, only fractions of the datasets were used. In particular, a total of 578 samples and 127 features comprised the AGP data set, while 500 samples and 26 features comprised the GG data set. In both data sets, 50% of the samples correspond to cases.

Validation scenarios results

Datasets were preprocessed as described in section "Data pretreatment". Further, the proposed method, as well as the SVM and SVM-RFE methods, were applied after separating the whole data set into training, testing, and validation sets using 70, 20, and 10% of the data, respectively. For the SVM-RFE method, the number of features to select was $n_{features} = \{5, 10, 15, \frac{n_{features}}{2}\}$ and the average of the results was calculated. The tuning parameters used for the SVM and SVM-RFE methods were $C = 1$ and $\gamma = 0.05$, where C trades off the correct classification of training examples against the maximization of the decision function's margin, and γ defines how far the influence of a single training example reaches.

Table 4 shows the main results. For the AGP data set, SVM is the least accurate, and SVM-RFE has the highest accuracy. This latter result is mostly due to all the strong features of SVM and the ability of the SVM-RFE method to eliminate variables that are not highly relevant in the data. Interestingly, our method outperforms SVM and is a close competitor of SVM-RFE.

For the GG dataset, although the number of variables is small, the SVM-RFE and our method showed accuracy values above 90%, while the accuracy for the SVM method is below this threshold. It is worth highlighting that, for this data set, our method outperforms both the SVM and SVM-RFE methods. The latter result is thanks to the artificial dimensional reduction conducted to balance the natural weights when the number of

Table 3 Performance of the proposed method for synthetic datasets. Configurations (n, m) not reported showed 100% Classification Accuracy

Sample size (n)	Number of features (m)	Classification Accuracy (%)
80	40	99.8
100	20	98.1
120	20	99.7
160	20	98.0

Table 4 Classification accuracy for each method for the AGP and GG data sets

Dataset	SVM	SVM-RFE	Proposed Method
AGP	92.03%	96.33%	95.06%
GG	89.34%	92%	94%

samples is greater than the number of variables. Figure 5 provides a graphical illustration of the proposed method's classification outcome for both real datasets used for validation, i.e., the AGP and the GG.

Discussion and conclusions

The ability to characterize populations of patients, species, or biological features, usually comprising a large number of variables in order to use the extracted characteristics to classify new samples into one of such populations' categories is a relevant tool for biological and medical studies. When data describing these populations is compositional, further limitations and challenges arise.

Here, we proposed a new method to classify samples into one of two previously known categories. The method uses a new metric developed to quantify the overall correlation structure deviation between two datasets, and a new dimensionality reduction technique. Although we illustrated the usefulness of our proposal with compositional data, its application is not limited, under any circumstances, to data of this nature. In fact, when data is not compositional, the centered log-ratio transformation and the zero-replacement algorithm must not be applied.

Validation with synthetic data showed that the proposed method achieves accuracy values above 98%. Moreover, comparison of the performance of our method with that of SVM and the SVM-RFE (i.e., two state-of-the-art classification techniques), using two real-world datasets from 16 s RNA sequencing experiments, showed that our method outperforms the SVM method in both data sets, outperforms the SVM-RFE method in the GG data set, and is a close competitor of the SVM-RFE method in the AGP data set.

Future studies may address the ability of our proposed method to perform accurately for a broader range of dimensions (number of variables and samples) and assess its performance for more scenarios of dissimilar correlation structures other than that for $\rho_c = 0.1$ and $\rho_v = 0.2$. Moreover, our method may be extrapolated for multi-category classification, and a performance assessment may be conducted to test its classification accuracy in non-binary scenarios.

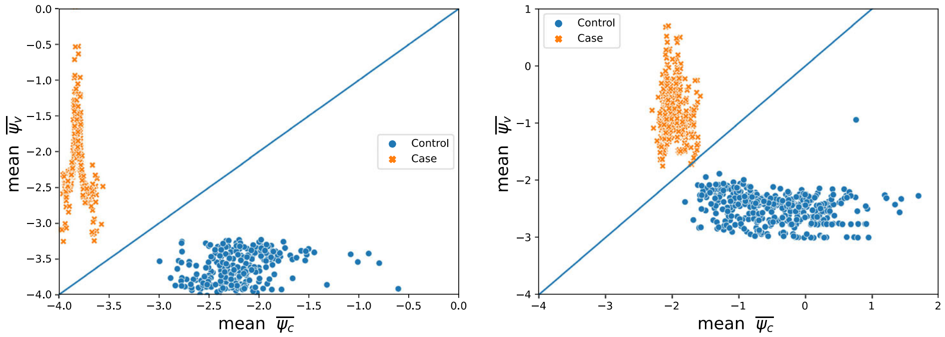

Fig. 5 Illustration of new samples and the line that separates both groups with the proposed method for the AGP (left) and GG (right) data sets

Abbreviations

AGP: American gut project; alr: Additive lo-ratio; ANN: Artificial Neural Networks; BM: Bayesian multiplicative (algorithm); clr: Centered log-ratio; CSVM-RMFE: Correlation based support vector machine–recursive multiple feature elimination; GG: Greengenes (database); GrBoost: Gradient boosting; ilr: Isometric log-ratio; NN: Nearest Neighbor; OTU: Operational taxonomic unit; PCA: Principal component analysis; rRNA: Ribosomal ribonucleic acid; SPIEC-EASI: Sparse inverse covariance estimation for ecological association inference; SVM: Support vector machine; SVM-RFE: Support vector machine recursive feature elimination

Acknowledgements and funding

This study was financed by COLCIENCIAS grant No. 1215-5693-4635, contract 0770-2013. By the time this work was developed, IP was a doctoral student at Universidad del Norte, Colombia, whose PhD was funded by COLCIENCIAS and Gobernación del Atlántico (Colombia), grant No. 673 (2014), "*Formación de Capital Humano de Alto Nivel para el Departamento del Atlántico*". EZ and HSJV are supported in part by award No. R01AI110385 from the National Institute of Allergy and Infectious Diseases, National Institutes of Health, Bethesda, MD, USA. JIV was partially supported by research grant FOFICO 32101 PE0031 from Universidad del Norte, Barranquilla, Colombia.

Authors' contributions

Technique design: SR, IP, EZ. Algorithms implementation: SR, IP. Experimental design: JIV, EZ, HSJV, MS. Writing of the manuscript: SR, IP, EZ, JIV, MS, HSJV.

Availability of data and materials

The source code, implemented in Python 3, is readily available in the following GitHub site: https://github.com/JoaoRacedo/arn_seq_pipeline. This code generates synthetic datasets to demonstrate the use of the pipeline. The American Gut Project's datasets can be found on the following website: http://americangut.org. Finally, the Greengenes' datasets can be found on: https://greengenes.lbl.gov/Download/OTUs/.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Received: 4 November 2020 Accepted: 16 June 2021
Published online: 09 July 2021

References

1. Turnbaugh PJ, Ley RE, Hamady M, Fraser-Liggett CM, Knight R, Gordon JI. The Human Microbiome Project. Nature [Internet]. 2007;449(7164):804–10. Available from: https://doi.org/10.1038/nature06244.
2. Kitano H. Looking beyond the details: a rise in system-oriented approaches in genetics and molecular biology. Curr Genet [Internet]. 2002 [cited 2019 Nov 13];41(1):1–10. Available from: https://doi.org/10.1007/s00294-002-0285-z.
3. Oltvai ZN. Life's complexity pyramid Zoltán N. Oltvai. 2010;763(2002).
4. Kitano H. Systems biology: a brief overview. 2015;(April 2002).
5. Voorhies AA, Ott CM, Mehta S, Pierson DL, Crucian BE, Feiveson A, et al. Study of the impact of long-duration space missions at the International Space Station on the astronaut microbiome. Sci Rep [Internet]. 2019;1–17. Available from: https://doi.org/10.1038/s41598-019-46303-8
6. Somerville C, Somerville S. Plant functional genomics. Science. 1999;285(5426):380–3.
7. Gill R, Datta S, Datta S. A statistical framework for differential network analysis from microarray data. BMC Bioinformatics. 2010;11(1):95.
8. Gill R, Datta S, Datta S. dna: an R package for differential network analysis. Bioinformation. 2014;10(4):233.
9. Juric D, Lacayo NJ, Ramsey MC, Racevskis J, Wiernik PH, Rowe JM, et al. Differential gene expression patterns and interaction networks in BCR-ABL—positive and—negative adult acute lymphoblastic leukemias. J Clin Oncol. 2007;25(11):1341–9.
10. Van Treuren W, Ren B, Gevers D, Kugathasan S, Denson LA, Va Y, et al. Resource the treatment-naive microbiome in new-onset Crohn's disease. Cell Host Microbe. 2014;15:382–92.
11. Ruan D, Young A, Montana G. Differential analysis of biological networks. BMC Bioinformatics. 2015;16(1):327.
12. Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, et al. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. Appl Environ Microbiol. 2009;75(23):7537–41.
13. Rao KR, Lakshminarayanan S. Partial correlation based variable selection approach for multivariate data classification methods. Chemom Intell Lab Syst. 2007;86(1):68–81.
14. Kurtz ZD, Müller CL, Miraldi ER, Littman DR, Blaser MJ, Bonneau RA. Sparse and compositionally robust inference of microbial ecological networks. PLoS Comput Biol. 2015;11(5):e1004226.
15. Aitchison J. The statistical analysis of compositional data. J R Stat Soc Ser B. 1982:139–77.
16. Filzmoser P, Hron K, Reimann C. Science of the Total Environment Univariate statistical analysis of environmental (compositional) data: problems and possibilities. Sci Total Environ [Internet]. 2009;407(23):6100–8. Available from: https://doi.org/10.1016/j.scitotenv.2009.08.008.

17. Clark C, Kalita J. A comparison of algorithms for the pairwise alignment of biological networks. Bioinformatics [Internet]. 2014;30(16):2351–9. Available from: https://doi.org/10.1093/bioinformatics/btu307.
18. Atchison J, Shen SM. Logistic-normal distributions: some properties and uses. Biometrika. 1980;67(2):261–72.
19. Aitchison J. A new approach to null correlations of proportions. J Int Assoc Math Geol. 1981;13(2):175–89.
20. Egozcue JJ, Pawlowsky-Glahn V, Mateu-Figueras G, Barceló-Vidal C. Isometric Logratio transformations for compositional data analysis. Math Geol [Internet]. 2003;35(3):279–300. Available from: https://doi.org/10.1023/A:1023818214614.
21. Greenacre M, Grunsky E. The isometric logratio transformation in compositional data analysis: a practical evaluation. 2019.
22. Pan M, Zhang J. Correlation-based linear discriminant classification for gene expression data. Genet Mol Res. 2017;16(1).
23. Hira ZM, Gillies DF. A review of feature selection and feature extraction methods applied on microarray data. Adv Bioinforma 2015;2015.
24. Goswami S, Chakrabarti A, Chakraborty B. Analysis of correlation structure of data set for efficient pattern classification. In: 2015 IEEE 2nd International Conference on Cybernetics (CYBCONF); 2015. p. 24–9.
25. Russell EL, Chiang LH, Braatz RD. Data-driven methods for fault detection and diagnosis in chemical processes. New York: Springer Science & Business Media; 2012.
26. Saeys Y, Inza I, Larrañaga P. A review of feature selection techniques in bioinformatics. Bioinformatics. 2007;23(19):2507–17.
27. Serban N, Critchley-Thorne R, Lee P, Holmes S. Gene expression network analysis and applications to immunology. Bioinformatics. 2007;23(7):850–8.
28. Friedman J, Alm EJ. Inferring correlation networks from genomic survey data. PLoS Comput Biol. 2012;8(9):e1002687.
29. Radovic M, Ghalwash M, Filipovic N, Obradovic Z. Minimum redundancy maximum relevance feature selection approach for temporal gene expression data. BMC Bioinformatics. 2017;18(1):1–14.
30. Anders S, Huber W. Differential expression analysis for sequence count data. Genome Biol. 2010;11(10):R106.
31. Paulson JN, Stine OC, Bravo HC, Pop M. Differential abundance analysis for microbial marker-gene surveys. Nat Methods. 2013;10(12):1200–2.
32. Kavitha KR, Rajendran GS, Varsha J. A correlation based SVM-recursive multiple feature elimination classifier for breast cancer disease using microarray. In: 2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI); 2016. p. 2677–83.
33. Collins GS, Mallett S, Omar O, Yu L-M. Developing risk prediction models for type 2 diabetes: a systematic review of methodology and reporting. BMC Med. 2011;9(1):103.
34. Aarøe J, Lindahl T, Dumeaux V, Sæbø S, Tobin D, Hagen N, et al. Gene expression profiling of peripheral blood cells for early detection of breast cancer. Breast Cancer Res. 2010;12(1):R7.
35. Datta S. Classification of breast cancer versus normal samples from mass spectrometry profiles using linear discriminant analysis of important features selected by random forest. Stat Appl Genet Mol Biol. 2008;7(2).
36. Šonka M, Hlaváč V, Boyle R. Image processing, analysis, and machine vision. International Student Edition; 2008.
37. Dembélé D, Kastner P. Fold change rank ordering statistics: a new method for detecting differentially expressed genes. BMC Bioinformatics. 2014;15(1):14.
38. Bevilacqua V, Mastronardi G, Menolascina F, Paradiso A, Tommasi S. Genetic algorithms and artificial neural networks in microarray data analysis: a distributed approach. Eng Lett. 2006;13(4).
39. Ca DAV, Mc V. Gene expression data classification using support vector machine and mutual information-based gene selection. Proc Comput Sci. 2015;47:13–21.
40. van Dam S, Vosa U, van der Graaf A, Franke L, de Magalhaes JP. Gene co-expression analysis for functional classification and gene--disease predictions. Brief Bioinform. 2018;19(4):575–92.
41. Dudoit S, Fridlyand J, Speed TP. Comparison of discrimination methods for the classification of tumors using gene expression data. J Am Stat Assoc. 2002;97(457):77–87.
42. Bhuvaneswari V, et al. Classification of microarray gene expression data by gene combinations using fuzzy logic (MGC-FL). Int J Comput Sci Eng Appl. 2012;2(4):79.
43. Belciug S, Gorunescu F. Learning a single-hidden layer feedforward neural network using a rank correlation-based strategy with application to high dimensional gene expression and proteomic spectra datasets in cancer detection. J Biomed Inform. 2018;83:159–66.
44. Dettling M, Bühlmann P. Boosting for tumor classification with gene expression data. Bioinformatics. 2003;19(9):1061–9.
45. Friedman J, Hastie T, Tibshirani R, et al. Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). Ann Stat. 2000;28(2):337–407.
46. Freund Y, Schapire RE. A decision-theoretic generalization of on-line learning and an application to boosting. J Comput Syst Sci. 1997;55(1):119–39.
47. Fix E, Hodges Jr JL. Discriminatory analysis-nonparametric discrimination: small sample performance; 1952.
48. Breiman L, Friedman J, Stone CJ, Olshen RA. Classification and regression trees. Boca Raton, FL: CRC Press; 1984.
49. Martín-Fernández J-A, Hron K, Templ M, Filzmoser P, Palarea-Albaladejo J. Bayesian-multiplicative treatment of count zeros in compositional data sets. Stat Modelling. 2015;15(2):134–58.
50. Pearson K. Mathematical contributions to the theory of evolution—on a form of spurious correlation which may arise when indices are used in the measurement of organs. Proc R Soc Lond. 1897;60(359–367):489–98.
51. McDonald D, Hyde E, Debelius JW, Morton JT, Gonzalez A, Ackermann G, et al. American Gut: an open platform for citizen science microbiome research. Msystems. 2018;3(3):e00031–18.
52. DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K, et al. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Appl Environ Microbiol. 2006;72(7):5069–72.

Publisher’s Note