Supervised Variational Relevance Learning, An Analytic Geometric Feature Selection with Applications to Omic Datasets

Marcelo Boareto, Jonatas Cesar, Vitor B.P. Leite, and Nestor Caticha

Abstract—We introduce Supervised Variational Relevance Learning (Suvrel), a variational method to determine metric tensors to define distance based similarity in pattern classification, inspired in relevance learning. The variational method is applied to a cost function that penalizes large intraclass distances and favors small interclass distances. We find analytically the metric tensor that minimizes the cost function. Preprocessing the patterns by doing linear transformations using the metric tensor yields a dataset which can be more efficiently classified. We test our methods using publicly available datasets, for some standard classifiers. Among these datasets, two were tested by the MAQC-II project and, even without the use of further preprocessing, our results improve on their performance.

Index Terms—Suvrel, Relevance Learning, analytic metric learning, proteomics, metabolomics, genomics, feature selection, distance learning

1 Introduction

MANY of the current biomolecular technologies permit measuring a large number of variables, generating high-dimensional data, often involving thousands or even millions of features. Very often a large fraction of these features are redundant or irrelevant and in this case, feature selection techniques are of fundamental importance for constructing a predictive model. Furthermore, from a biological point of view, inspecting a shorter list of biomarkers candidates can shed light on the biological process involved. Other benefits include shorter training times and the reduction of overfitting.

The archetypal case is the use of feature selection applied to DNA microarrays data, where there are on the order of ten thousand genes, whereas only a small subset of these genes is relevant, and the number of experiments range from a few tens up to a few hundreds. Much effort has been invested in defining a good methodology for outcome prediction using microarray data [1]. It has been shown that feature selection methods have a significant influence in the accuracy of predictive tests using microarray data and surprisingly the simple Student's t-test seems to provide better results than sophisticated methods [2]. In addition, results such as those from the Microarray Quality Control (MAQC)-II study indicate that the choice of the number of features has a higher impact on the method performance than the feature selection method, see Fig. 5a of [1].

An essential ingredient of classification or clustering by similarity is to define an appropriate definition of similarity or alternatively, distance. Our central problem is to define a metric tensor which incorporates the statistical significance of each feature and hence defines the relevant geometrical space where the data is represented. We present a methodology for data preprocessing, Supervised Variational Relevance Learning (Suvrel), that extends ideas by [3], [4], [5], [6], [7] for relevance learning. This has also been studied under the name distance learning, e.g. see [8], [9], [10]. We suppose a scenario where a classifier will be used after preprocessing for feature selection by Suvrel has been performed. The feature selection preprocessing will hinge in determining an adequate metric tensor. The idea is to choose a metric such that patterns of the same class be represented closer to each other in comparison with elements of different groups. This is done by shrinking the non relevant dimensions and stretching the relevant ones accordingly to a measure of relevance. This new representation of the dataset can be useful to improve the classification accuracy of different methods and also for visualization of the data in low dimensions. The metric tensor can also be used to attribute statistical significant relevance to the features. But this relevance determination does not preclude using a further feature selection method which might benefit from an improved starting point. In a previous study, we applied related ideas to an enzyme classification problem [11]. The new result we present in this manuscript, is a variational method that can be solved exactly analytically. We applied this methodology to several datasets and our results indicate the advantages of representing the data in this new metric. The datasets include two sets obtained from MAQC-II project [1] that can be divided according to two clinical endpoints. For cases of more than two classes, we used a microarray leukemia dataset and a metabolite dataset.

- *M. Boareto, J. Cesar, and N. Caticha are with the Instituto de Física, University of São Paulo, São Paulo 05508, Brazil. E-mail: {marceloboareto, jonataseduardo}@gmail.com, nestor@if.usp.br.*
- *V.B.P. Leite is with the IBILCE, Universidade Estadual Paulista, São José do Rio Preto, São Paulo, Brazil. E-mail: vleite@sjrp.unesp.br.*

Manuscript received 3 Apr. 2014; revised 30 July 2014; accepted 17 Nov. 2014. Date of publication 4 Dec. 2014; date of current version 1 June 2015.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TCBB.2014.2377750

2 Methods

A datum from an experiment i is a vector $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{iF})$ in the space of features $\mathbb{R}^F$, where each feature μ represents a dimension (with $\mu = 1, \ldots, F$). The learning and validation sets, $\mathcal{L}$ and $\mathcal{V}$, are sets of pairs $\{(\boldsymbol{x}_i, c_i)\}$, where data vectors $\boldsymbol{x}_i$ come together with its class label c_i. The set of class labels values that c_i can take is $\mathcal{C} = \{c^1, c^2, \ldots, c^K\}$, where K is the total number of classes. The number of samples of each class is $n_{c^1}, n_{c^2}, \ldots, n_{c^K}$ and $n = \sum_{a \in \mathcal{C}} n_a$ is the total number of elements of a set.

The geometry of the space of features is not established until a measure of similarity is specified. We choose to define the underlying geometry by imposing a structure through a metric tensor $g_{\mu\nu}$. This idea goes back to Gauss and Riemann and has entered Physics via ideas in gravitation. In the Computer Science and Statistics literature, where more general types of similarity measures are used, the Riemann distance inspired the linear mapping called the Mahalanobis distance [12]. The relevant information to determine the metric structure, to be obtained by a variational process, is the class membership of the vectors of the learning set. Features will be considered relevant if they contribute to decrease intraclass or increase interclass average distances. If we restrict the analysis to the simpler case of a diagonal metric tensor, then each dimension μ (e.g. gene in case of microarrays datasets) has a relevance weight w_μ, the component of the diagonal metric tensor, still to be determined by minimizing a cost function. The distance between two elements i and j is defined as

$$d_{ij}^2 = \sum_{\mu\nu} g_{\mu\nu} \Delta x_{ij\mu} \Delta x_{ij\nu}, \tag{1}$$

where $\Delta x_{ij\nu} = x_{i\nu} - x_{j\nu}$ is the difference for the νth feature of the patterns i and j.

Cost functions that penalizes features that increase intraclass distances while favoring features that decrease interclass distances have been considered before, giving rise to a large body of literature which has been recently reviewed in [9], [10]. Our main contribution is that the specific form we consider is amenable to analytic manipulation so that the metric tensor can be obtained explicitly and that it can be established that the method renders a positive definite metric tensor. The cost function to be minimized with respect to the metric tensor, is

$$E(\{g_{\mu\nu}\}; \mathcal{L}) = \sum_{a \in \mathcal{C}} \left\langle d_{ij}^2 \right\rangle_{a,a} - \gamma \sum_{<a \neq b> \in \mathcal{C}} \left\langle d_{ij}^2 \right\rangle_{a,b}, \tag{2}$$

where a and b refer to class index, and γ a parameter that controls the weight of the interclass distance term, with respect to the intraclass distance term. The notation $<a \neq b>$ denotes that different class pairs are included only once. The angle brackets denote an average, given by

$$\left\langle d_{ij}^2 \right\rangle_{a,b} = \frac{1}{n_{ab}} \sum_{i \in a, j \in b} d_{ij}^2, \tag{3}$$

where $n_{ab} = n_a n_b$ and $\sum_{i \in a}$ is the summation over all samples that belong to class a. Since the distance defined in (1) depends on a summation over the features, we can rewrite the cost function (2) as

$$E = \sum_{\mu\nu} g_{\mu\nu} \epsilon_{\mu\nu}, \tag{4}$$

where $\epsilon_{\mu\nu}$ is given by

$$\epsilon_{\mu\nu} = e_{\mu\nu}^{in} - \gamma e_{\mu\nu}^{out} \tag{5}$$

and

$$\begin{aligned} e_{\mu\nu}^{in} &= \sum_{a \in \mathcal{C}} \frac{1}{n_{aa}} \sum_{i,j \in a} \Delta x_{ij\mu} \Delta x_{ij\nu}, \\ &= \sum_{a \in \mathcal{C}} \frac{1}{n_a^2} \left(n_a \sum_{i \in a} x_{i\mu} x_{i\nu} + n_a \sum_{j \in a} x_{j\mu} x_{j\nu} \right. \\ &\quad \left. -2 \sum_{i \in a} x_{i\mu} \sum_{j \in a} x_{j\nu} \right) \\ &= 2 \sum_{a \in \mathcal{C}} \mathrm{cov}(\boldsymbol{x}_\mu^a; \boldsymbol{x}_\nu^a) \end{aligned} \tag{6}$$

$$\begin{aligned} e_{\mu\nu}^{out} &= \sum_{<a \neq b> \in \mathcal{C}} \frac{1}{n_{ab}} \sum_{i \in a, j \in b} \Delta x_{ij\mu} \Delta x_{ij\nu}, \\ &= \sum_{<a \neq b> \in \mathcal{C}} \frac{1}{n_a n_b} \left(n_b \sum_{i \in a} x_{i\mu} x_{i\nu} + n_a \sum_{j \in b} x_{j\mu} x_{j\nu} \right. \\ &\quad \left. - \sum_{i \in a} x_{i\mu} \sum_{j \in b} x_{j\nu} - \sum_{i \in a} x_{i\nu} \sum_{j \in b} x_{j\mu} \right) \\ &= \sum_{<a \neq b> \in \mathcal{C}} \left[\mathrm{cov}\left(\boldsymbol{x}_\mu^a; \boldsymbol{x}_\nu^a\right) + \mathrm{cov}\left(\boldsymbol{x}_\mu^b; \boldsymbol{x}_\nu^b\right) \right. \\ &\quad \left. + (m_{a\mu} - m_{b\mu})(m_{a\nu} - m_{b\nu}) \right]. \\ &= (K-1) \sum_{a \in \mathcal{C}} \mathrm{cov}\left(\boldsymbol{x}_\mu^a; \boldsymbol{x}_\nu^a\right) \\ &\quad + \sum_{<a \neq b> \in \mathcal{C}} (m_{a\mu} - m_{b\mu})(m_{a\nu} - m_{b\nu}). \\ &= (K-1) \sum_{a \in \mathcal{C}} \mathrm{cov}\left(\boldsymbol{x}_\mu^a; \boldsymbol{x}_\nu^a\right) + K^2 \mathrm{cov}(\boldsymbol{m}_\mu; \boldsymbol{m}_\nu), \end{aligned} \tag{7}$$

where $\boldsymbol{x}_\mu^a$ is the set of samples of the μ feature that belongs to the class a and $m_{a\nu} = \frac{1}{n_a} \sum_{i \in a} x_{i\nu}$. The first term has the sum over classes of the covariance within a class of the features. The second term is the covariance of the vector of dimension K whose components are the means over a class: $\boldsymbol{m}_\mu = \{m_{a\mu}\}$,

$$\mathrm{cov}(\boldsymbol{m}_\mu; \boldsymbol{m}_\nu) = \frac{1}{K} \sum_{a \in \mathcal{C}} m_{a\mu} m_{a\nu} - \left(\frac{1}{K} \sum_{a \in \mathcal{C}} m_{a\mu} \right) \left(\frac{1}{K} \sum_{a \in \mathcal{C}} m_{a\nu} \right).$$

It follows that

$$\begin{aligned} \epsilon_{\mu\nu} &= [2 - (K-1)\gamma] \sum_{a \in \mathcal{C}} \mathrm{cov}\left(\boldsymbol{x}_\mu^a; \boldsymbol{x}_\nu^a\right) \\ &\quad - \gamma K^2 \mathrm{cov}(\boldsymbol{m}_\mu; \boldsymbol{m}_\nu), \end{aligned} \tag{8}$$

which only depends on the data vectors and on the supervising information of which experiment i belongs to each class a.

Relevance is obtained by imposing that E is a minimum with respect to the metric tensor variations subject to a scale constraint $\sum_{\mu\nu} g_{\mu\nu}^2 = 1$, with γ fixed. The solution, obtained using the standard method of Lagrange multipliers

$$\frac{\delta}{\delta g_{\mu\nu}} \left(E + \theta \left(\sum_{\mu\nu} g_{\mu\nu}^2 - 1 \right) \right) = 0 \tag{9}$$

is

$$g_{\mu\nu} = \frac{-\epsilon_{\mu\nu}}{\sqrt{\sum_{\mu'\nu'} \epsilon_{\mu'\nu'}^2}}. \tag{10}$$

Once we obtain the metric tensor $g_{\mu\nu}$, we can define a distance between the data points using Eq. (1) or alternatively apply a coordinate transformation with the square root of the new metric tensor, but for this we have to investigate conditions that lead to a positive definite metric tensor. This is discussed in Appendix 5.1 which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TCBB.2014.2377750, where we prove that there is a $\gamma^* < 2/(K-1)$ such that for $\gamma > \gamma^*$ the metric tensor is positive definite. For the diagonal case, with

$$d_{ij}^2 = \sum_{\nu} w_\mu (x_{i\nu} - x_{j\nu})^2, \tag{11}$$

the individual feature cost ϵ_ν can be written in terms of the averages and variances $\sigma_{a\nu}^2 = \frac{1}{n_a} \sum_{i \in a} x_{i\nu}^2 - m_{a\nu}^2$

$$\epsilon_\nu = (2 - (K-1)\gamma) \sum_a \sigma_{a\nu}^2 - \gamma \sum_{a \neq b} (m_{a\nu} - m_{b\nu})^2, \tag{12}$$

which is the form we will use in the numerical applications.

For the case of two classes ($K = 2$) the previous equation reduces to

$$\epsilon_\nu = (2 - \gamma) \sum_a \sigma_{a\nu}^2 - \gamma \sum_{a \neq b} (m_{a\nu} - m_{b\nu})^2. \tag{13}$$

Then $\gamma = 2$ is an acceptable value for classification purposes. If $\gamma < \gamma^*$ the metric tensor can assume negative values. On the other hand, if $\gamma > 2$ the term involving the intraclass variance becomes negative, which means that the increase of intraclass variance decreases the cost, hampering distinguishability among the classes.

For $\gamma = 2$ we have

$$\epsilon_\nu = -2 \sum_{a \neq b} (m_{a\nu} - m_{b\nu})^2, \tag{14}$$

so that the metric component for feature ν is proportional to the difference of the averages between the conditions. For the case of two classes, a suitable normalization leads to the following relation:

$$\epsilon_\nu = -2 t_\nu^2, \tag{15}$$

where t_ν is the Student's t-variable. Then, up to an overall scale each component of the relevance vector $\boldsymbol{w}$ is proportional to the square of the t-variable and consequently to the p-value of a t-test, given by,

$$\omega_\nu = \frac{t_\nu^2}{\sqrt{\sum_{\nu'} t_{\nu'}^4}}. \tag{16}$$

We now show results for the diagonal case and address in Appendix 5.2 in the supplemental materials, available online, how to treat off-diagonal cases.

3 Results

An important application of this method is to use it in conjunction with a classification procedure, by preprocessing the training and validation datasets. A measure of its usefulness comes from

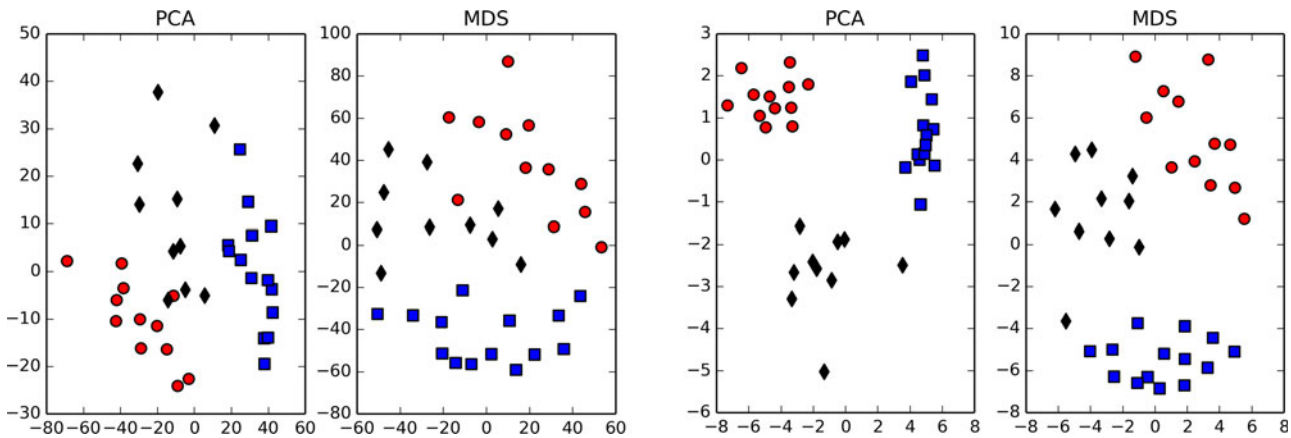

Fig. 1. Two dimensional representation of leukemia dataset using Principal Component Analysis and Multidimensional Scaling (Left) Same relevance assigned to all genes. (Right) Relevance assigned by Suvrel. The symbols for the ALL class are red circles, for AML are blue squares and for MLL are black diamonds.

comparing the performances of a classifier with and without the preprocessing.

For classification into two classes, where $\chi = a$ or b is the prediction and $\tau = a$ or b is the true class, a measure of the prediction performance is given by Matthews correlation coefficient, which we write in a convenient form as

$$MCC = \sqrt{\frac{n_a^\tau n_b^\tau}{n_a^\chi n_b^\chi}} \det \boldsymbol{P}, \tag{17}$$

where the matrix $\boldsymbol{P}$ has elements $P_{\chi\tau} = P(\chi|\tau)$; and n_a^τ and n_a^χ are respectively the correct and the predicted number of vectors in class a. This is a measure that takes into account true and false positives/negatives in a balanced way [13], [14]. When MCC is equal to one it means that the method classified correctly all samples, conversely, if MCC is equal to minus one it means that classification was completely incorrect. The main reason for using this figure of merit is that our results can be compared with the MicroArray Quality Control consortium. We studied the performance of several variations of k-Nearest Neighbor (KNN) and Support Vector Machines (SVM). For the datasets where there are more than two classes, leukemia (3 classes) and cow diet (4 classes), we used a generalization of the MCC discussed in [14] and compared our methodology with feature selection using ANOVA. All raw microarray data were preprocessed using the RMA method [15]. Codes are available and their location can be found in Appendix 5.3 in the supplemental materials, available online.

3.1 MAQC Consortium Datasets

We first applied our methodology to two microarray datasets obtained from the MicroArray Quality Control consortium [1]. Each dataset consists of a training and validation set generated by different experimental groups. The Breast Cancer dataset can be divided according to two endpoints: pre-operative treatment response (pCR) and estrogen receptor (ER). The Multiple Myeloma dataset can be divided according to overall survival milestone outcome (OS-MO), event free survival milestone outcome (EFS-MO) and a "negative" endpoint where the classes are chosen randomly (CPR1). For more details about the class distribution see Table 1. For these cases of a two-class classification, we choose $\gamma = 2$ and a suitable normalization so that the metric tensor is proportional to the Student t-variable.

We compared our approach with different feature selection methods based on the t-test, which are: the top 10 and top 100 genes ranked according to its p-value, and using the leave one out (LOO) procedure in the training dataset in order to select the best number of genes. The case were all genes are considered to have the same relevance is also compared. In the majority of the cases, the best performance was obtained using the Suvrel preprocessing, see Tables 3, 4, 5 and 6. We also compared our results with 17 MAQC data analysis teams that nominated their best models for microarray datasets [1]. In the majority of the cases our approach is superior to the median of the teams performance and frequently is higher than the best MAQC candidate. These results are summarized in Figs. 3 and 4 which also present a comparison of the prediction performance for the case without preprocessing and with Suvrel, and a comparison of Suvrel and other algorithms, respectively.

We also used the negative control generated by MAQC for the multiple myeloma dataset. In this case, the classes are assigned randomly and as expected all methodologies have a MCC close to zero, see Table 7.

3.2 Leukemia Dataset

We also applied our methodology to a leukemia dataset that consists of 24 experiments with patients with Acute Lymphoblastic Leukemia (ALL), 28 experiments with patients with Acute Myelogenous Leukemia (AML) and 20 experiments with patients with Mixed-Lineage Leukemia (MLL) [16]. The dataset was divided in training and validation sets according to the methodology described in the original manuscript [16]. The data was normalized by subtracting the average and dividing by the standard deviation of each gene, zero mean and unit variance normalization.

Using the training dataset, we applied our methodology in order to determine the metric vector. Then, the distribution of the points for the validation dataset was represented in two dimensions using Principal Component Analysis (PCA) and Multidimensional Scaling (MDS) [17], Fig. 1. Visual inspection shows a much better result, that can be quantified by noticing that for the first two PCA eigenvectors the explained variance ratio goes from $[0.346, 0.078]$ without preprocessing to $[0.493, 0.096]$ with Suvrel. In order to quantify the change in prediction performance, we compared our approach with a feature selection procedure based on ANOVA, using genes with a p-value lower than 0.05. Suvrel presents a superior performance for all classification methods, see Figs. 3, 4 and Table 8.

3.3 Cow Diet Metabolites Dataset

The last set consists of data from a metabolite experiment where the concentrations of 39 rumen samples was measured by proton NMR from dairy cows fed with different proportions of barley grain. The groups are divided in four classes: 0, 15, 30, or 45 percent indicating the percentage of grain in the diet [18].

Again using PCA and MDS the dataset is represented in two dimensions for both cases with no feature selection and Suvrel (Fig. 2). A better class distinction is found in the new metric, mainly for the experiments with 30 and 45 percent of grain in diet. For the

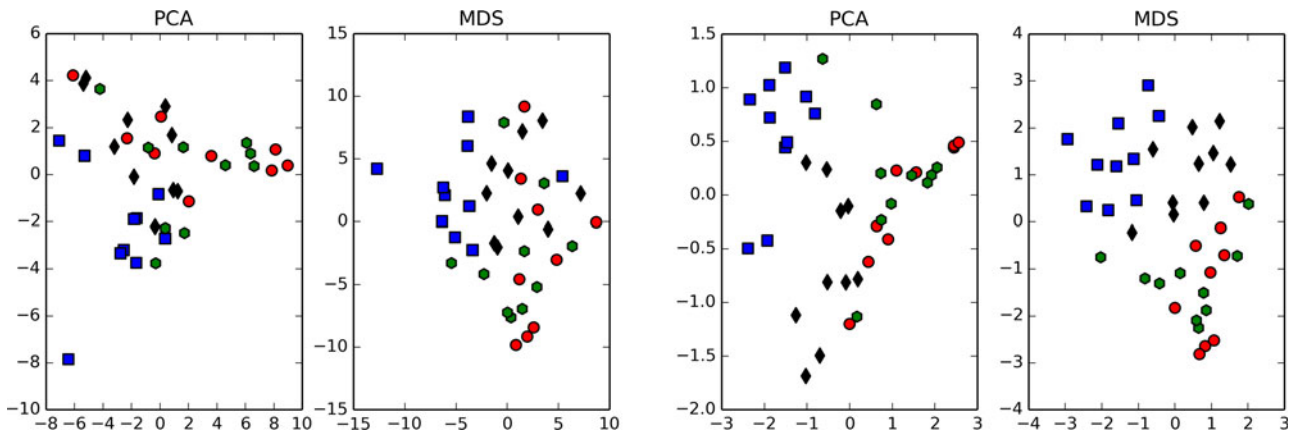

Fig. 2. Two dimensional representation of cow diet metabolites dataset using Principal Component Analysis and Multidimensional Scaling. (Left) Same relevance assigned to all genes. (Right) Relevance assigned by Suvrel. (red circles) 0 percent, (green hexagons) 15 percent, (black diamonds) 30 percent, (blue squares) 45 percent of grain in cow diet.

first two PCA eigenvectors the explained variance ratio goes from $[0.358, 0.134]$ without preprocessing to $[0.469, 0.132]$ with Suvrel. We also observed a good improvement in class prediction, Table 2. In this case, since there is a small number of experiments, we used the leave one out methodology to determine the accuracy of the methods. See the comparison in Figs. 3 and 4.

4 Discussion and Conclusions

We have introduced a new analytic method for feature selection that is based on a variational method applied to cost functions that include intra and interclass information and that can be applied to any number of classes. The variational procedure is simple enough that an analytical solution is available. A parameter γ weights the relative importance given to the cost associated to intraclass and interclass distances. We have shown that the metric tensor is positive definite in a region around $\gamma = 2/(K-1)$ where K represents the number of classes. Then, in the case of two-class classification ($K = 2$), the diagonal of the metric tensor, which can be interpreted as measuring feature relevance, is proportional to the square of Student's t-variable. Ranking the features according to a statistical test and then selecting a restricted number of features is a standard approach of feature selection. However, choosing the number of predictive features is a question that has to be carefully addressed and it is usually done by maximizing the classification performance on the training dataset, with respect to the number of features. This procedure is very reasonable if the training and validation datasets are consistent, in the sense that interchanging learning and validations sets is of no great consequence. This is not the case of microarrays where the lack of robustness between the predictive lists of different datasets is a well known issue [19], [20]. In this case, the prediction model becomes very sensitive to the number of features chosen, as highlighted by the MAQC-II consortium [1]. Our method does not impose selection of a fixed number of features, which might explain why it showed a good improvement in the accuracy when compared to the standard methodology despite being based on the same ranking measure, which is the Student's t-variable. Combining Suvrel with standard classifiers (KNN and SVM) turns out to be, often, superior to the well elaborated protocols used by the MAQC-II teams (Fig. 3), therefore suggesting that this is a very useful methodology for feature section.

Over the past years, metabolomics analysis has become an important tool for understanding biological processes of interest. However, there is a lack of methodologies for selecting the lists of relevant metabolites. Usually, the data points are represented in a lower space by PCA and if this representation shows a good

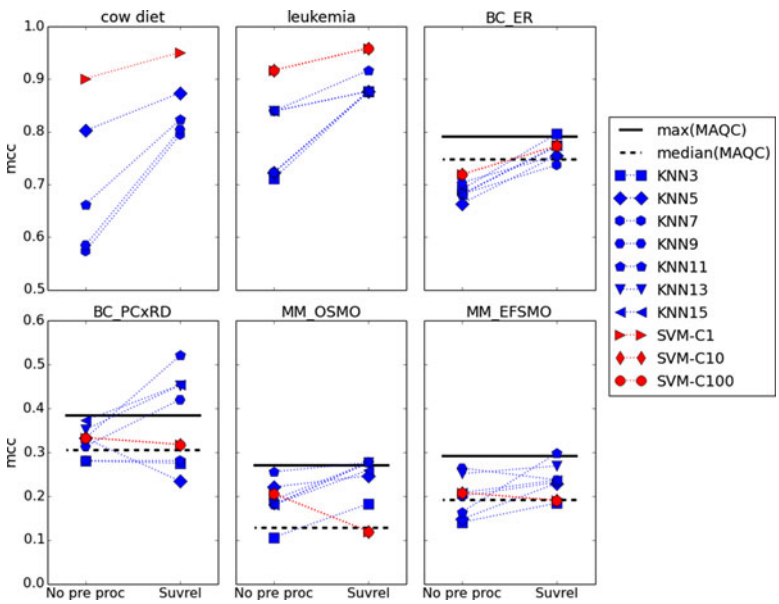

Fig. 3. Each panel shows the results for each of the datasets of the prediction accuracy for (left) no data preprocessing, e.g. same relevance assigned to all genes, and (right) the case where the data is preprocessed by Suvrel. The horizontal lines indicate the median and maximum performance of MAQC teams.

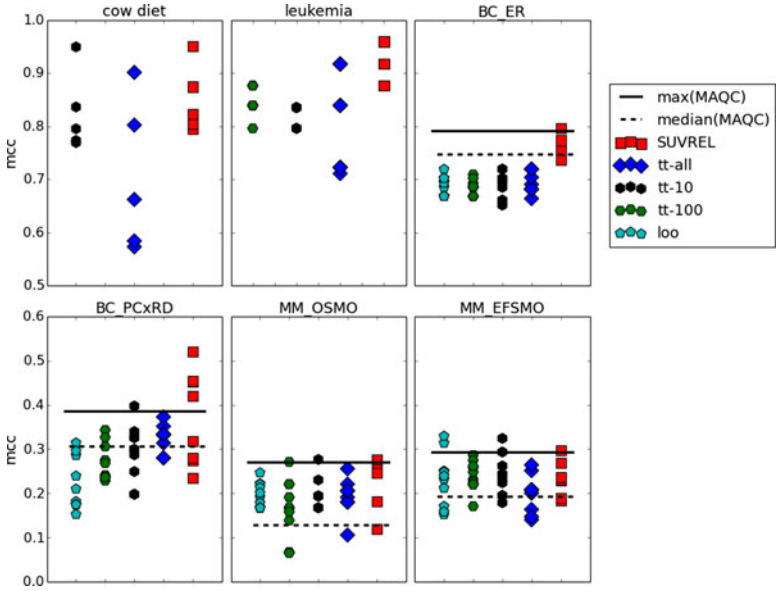

Fig. 4. Each panel shows the results for each of the datasets of the prediction accuracy for different preprocessing methods using different versions of KNN and SVM for classification. The horizontal lines indicate the median and maximum performance of MAQC teams.

separation between the classes, the metabolites that represent the dimensions that capture the higher variance are considered relevant. Furthermore, we point out that this might not be the best procedure and to chose the dimensions that contribute the most to the clusterization of the groups should be a better choice. Then, Suvrel can be a useful tool for selecting the relevant metabolites.

Due to the high dimensional nature of many bioinformatics data, the use of feature selection techniques has become an indispensable tool for biological model building. We point out that, unlike the majority of metric learning algorithms - with require a numerical minimization algorithm - Suvrel offers an analytical solution to find the relevant set of features. This provides an automatic and computational cheap method that can be useful for large datasets. Furthermore, the underlining assumptions of the method- presented in the form of a cost function - are straightforward to understand and the resulting analytical solution of the metric vector can be interpreted in terms of differences in averages and variances and even, for a particular value of γ, in terms of the well known t-Student variable. Among the extensions to the method we used in the numerical part of this paper we mention two. First we can consider the interactions among the features, which is done by taking the full non-diagonal metric tensor into account. This seems to be an immediate logical extension, since correlation of features is an intrinsic aspect of many biological systems. For instance, in gene expression often the determination of a given class is associated with a poll of features. In this case, it is expected an occurrence of significant (non-diagonal) cross-correlation terms. Second, a very simple online method can be used to update the full metric tensor, to accommodate the classification of new patterns, will be published elsewhere.

5 Supplementary Material

5.1 Positivity of the Metric Tensor

Let u, v be arbitrary unit length vectors in the space of features. Define $Z^2(u) = \sum_a Z_a^2$ where $Z_a = \sum_{\mu, i\in a}(x_{i\mu} - m_{a\mu})u_\mu$ and $W^2(u) = \sum_{<a\neq b>} W_{ab}^2$ where $W_{ab} = \sum_\mu (m_{a\mu} - m_{b\mu})u_\mu$. Introduce

$$Q(u, v) = (2 - (K-1)\gamma)Z^2(u) - \gamma W^2(v). \tag{18}$$

Note that for $u = v$, $Q(u, u) = \sum_{\mu\nu}\epsilon_{\mu\nu}u_\mu u_\nu$ is a quadratic form. The condition for positivity of the metric tensor is that $Q(u, u) < 0$ for any unit lenght u. For $\gamma > 2/(K-1)$ this is trivial. We now show that the proof of positivity can be extended to smaller values of γ. Call v_* the vector for which $W^2(v_*) = \min W^2 > 0$. Also call u^* the vector such that $Z^2(u^*) = \max Z^2 < \infty$. Then, if $\gamma < 2/(K-1)$ we have $Q(u, u) \leq Q(u^*, v_*)$. We also have that $Q(u^*, v_*) < 0$ if

$$\gamma > \gamma_* = \frac{2}{\frac{W^2(v_*)}{Z^2(u^*)} + (K-1)}. \tag{19}$$

and so $\gamma_* < 2/(K-1)$. Hence the metric tensor is positive definite if $\gamma > \gamma^*$. There is a range of values $\gamma^* < \gamma \leq 2/(K-1)$ where positive definiteness is preserved. The only condition we assumed necessary is that $W^2(v_*) > 0$, which is trivially satisfied if at least two classes have different means.

5.2 Off-Diagonal Terms

For the diagonal metric tensor case where $g_{\mu\nu} = \omega_\mu\delta_{\mu\nu}$ and with $\boldsymbol{h}^{(0)} = diag(\sqrt{\omega_1}, \ldots, \sqrt{\omega_F})$ we can apply the rescaling

$$\boldsymbol{x}' = \boldsymbol{h}^{(0)}\boldsymbol{x} \tag{20}$$

to the entire dataset. Now, standard methods can be applied to study the geometry of experiments and other statistical properties.

If the off diagonal terms are small but we want to still include them, we can proceed perturbatively in a simple manner as long as the metric tensor is a positive-semidefinite matrix. Call $\boldsymbol{h} = \{h_{\mu\nu}\}$ the principal square root, i.e. $g_{\mu\nu} = \sum_\rho h_{\mu\rho}h_{\rho\nu}$, and we separate into the diagonal part and the off diagonal $g_{\mu\nu} = \omega_\mu\delta_{\mu\nu} + \epsilon a_{\mu\nu}$, where ϵ is small. Then expanding $h_{\mu\nu}$:

$$h_{\mu\nu} = h_{\mu\nu}^{(0)} + \epsilon h_{\mu\nu}^{(1)} + \epsilon^2 h_{\mu\nu}^{(2)},$$

we obtain up to second order in ϵ

$$h_{\mu\nu}^{(0)} = \omega_\mu^{1/2}\delta_{\mu\nu} \tag{21}$$

$$h_{\mu\nu}^{(1)} = \frac{1}{\omega_\mu^{1/2} + \omega_\nu^{1/2}}a_{\mu\nu} \tag{22}$$

$$h_{\mu\nu}^{(2)} = -\frac{1}{\omega_\mu^{1/2} + \omega_\nu^{1/2}}\sum_\rho h_{\mu\rho}^{(1)}h_{\rho\nu}^{(1)} \tag{23}$$

and the rescaling is

$$\boldsymbol{x}' = \boldsymbol{h}\boldsymbol{x}. \tag{24}$$

The complexity of calculating this off diagonal rescaling grows with the cube of the number of features plus the inevitable quadratic dependence on the number of samples for an offline method. For lower dimensional problems $\boldsymbol{h}$ can be obtained via SVD. Note that if there is a negative eigenvalue it can only occur if $\gamma < \gamma^*$. From continuity of the eigenvalues in γ it will be the smallest in absolute value. SVD can be used to throw away non-relevant features combinations associated to negative eigenvalues and obtain a pseudo-square root matrix in the spirit of the Moore-Penrose pseudo-inverse for values of γ even lower than γ^*.

5.3 Codes

Code was written in Python. For the classifiers we used code from the SciKit-Learn project [17]. Code is available at the github repository: `http://www.github.com/jonataseduardo/suvrel`

5.4 Tables

TABLE 1
Class Distribution of Each Endpoint for the Breast Cancer and Multiple Myeloma Dataset

	Training set			Validation set		
Data set	NST	PT	NT	NSV	PV	NV
Breast Cancer (pCR)	130	33	97	100	15	85
Breast Cancer (ER)	130	80	50	100	61	39
Multiple Myeloma (OS-MO)	340	51	289	214	27	187
Multiple Myeloma (EFS-MO)	340	84	256	214	34	180
Multiple Myeloma (CPR1)	340	200	140	214	122	92

Both datasets were generated by the MAQC-II consortium [1]. Columns are : Name of Data set, NST: Number of samples in the training set, PT: number of positive labels in training set, NT: number of negative labels in training set, NSV: number of samples in validation set, PV: number of positive labels in validation set, NV: number of negative labels in validation set.

TABLE 2
Results for the Cow Diet Metabolites Dataset

Method	Suvrel	no selection	Anova ($p < 0.05$)
SVM	**0.951**	0.901	0.949
KNN3	0.874	0.802	0.836
KNN5	0.795	0.573	0.769
KNN7	0.804	0.584	0.773
KNN9	0.823	0.662	0.795

Performance measured by the MCC for classification using several KNN and SVM classifier with data preprocessed using feature selection methods: (i) Suvrel, (ii) no selection, all features are used, (iii) feature selection by Anova. Our methodology shows a superior performance for all combination of prediction methods. The best performance is highlighted in bold.

TABLE 3
Results for the Breast Cancer Dataset, Divided According to Estrogen Receptor Status

	Suvrel	all	top-10	top-100	LOO
KNN11	0.7735	0.6812	0.6955	0.6681	(0.695, 10)
KNN13	0.7735	0.6812	0.7191	0.6681	(0.719, 20)
KNN15	0.7735	0.6812	0.7191	0.6681	(0.685, 50)
KNN3	**0.7964**	0.6904	0.6517	0.6927	(0.702, 20)
KNN5	0.7550	0.6637	0.7018	0.7091	(0.693, 150)
KNN7	0.7369	0.6817	0.6607	0.6848	(0.702, 350)
KNN9	0.7550	0.7036	0.6607	0.7018	(0.668, 130)
SVM-C1	0.7735	0.7191	0.6517	0.6861	(0.668, 280)
SVM-C10	0.7735	0.7191	0.6848	0.6861	(0.668, 280)
SVM-C100	0.7735	0.7191	0.7018	0.6861	(0.668, 280)

Performance as measured by the MCC for classification using several KNN and SVM classifiers with data preprocessed using feature selection methods: (i) Suvrel, (ii) all features, (iii) top 10 (iv) top 100 features selected by the t-test and (v) the top features chosen using the leave one out procedure (MCC, # features). MAQC teams best performance $MCC = 0.792$ and median 0.748. Performances superior to MAQC teams are highlighted in bold. Notation: KNN3 = KNN with 3 neighbors. SVM-C1 = SVM with C = 1.

TABLE 4
Breast Cancer Dataset Divided According to Pre-Operative Treatment Response (pCR)

	Suvrel	all	tt-10	tt-100	LOO
KNN11	**0.5206**	0.3318	0.2994	0.3430	(0.3140, 600)
KNN13	**0.4530**	0.3514	0.3256	0.2287	(0.2392, 150)
KNN15	**0.4533**	0.3726	0.2994	0.3058	(0.2858, 150)
KNN3	0.2750	0.2801	0.2492	0.2412	(0.1817, 200)
KNN5	0.2341	0.3318	0.3256	0.2750	(0.1528, 670)
KNN7	0.2801	0.2801	0.3396	0.2684	(0.2100, 550)
KNN9	**0.4201**	0.3135	**0.3972**	0.3267	(0.2957, 630)
SVM-C1	0.3178	0.3335	0.2870	0.2362	(0.1746, 420)
SVM-C10	0.3178	0.3335	0.1979	0.2362	(0.1746, 420)
SVM-C100	0.3178	0.3335	0.1979	0.2362	(0.1746, 420)

Columns are the same as in Table 3. MAQC teams best performance $MCC = 0.385$ and median 0.306. Performances superior to MAQC teams are highlighted in bold.

TABLE 5
Multiple Myeloma Dataset Divided According to Overall Survival Milestone Outcome (OS-MO)

	Suvrel	tt-all	tt-10	tt-100	loo
KNN11	**0.2765**	0.2556	0.1941	0.2207	(0.2219, 380)
KNN13	**0.2765**	0.1803	0.1941	0.2207	(0.2466, 30)
KNN15	0.2593	0.1803	0.1941	0.1912	(0.2207, 60)
KNN3	0.1826	0.1060	0.2302	0.1390	(0.1665, 20)
KNN5	0.2466	0.2207	**0.2765**	0.2710	(0.2132, 30)
KNN7	**0.2765**	0.1912	0.1941	0.1675	(0.1912, 120)
KNN9	**0.2765**	0.1803	0.1941	0.1675	(0.2008, 30)
SVM-C1	0.1197	0.2058	0.1675	0.1585	(0.1783, 60)
SVM-C10	0.1197	0.2058	0.1675	0.0675	(0.1675, 10)
SVM-C100	0.1197	0.2058	0.1675	0.0653	(0.1675, 10)

Columns are the same as in Table 3. MAQC teams best performance $MCC = 0.271$ and median 0.129. Performances superior to MAQC teams are highlighted in bold.

TABLE 6
Multiple Myeloma Dataset Divided According to Event Free Survival Milestone Outcome (EFS-MO)

	Suvrel	all	tt-10	tt-100	loo
KNN11	**0.2978**	0.1637	0.2434	0.2206	(0.2317, 170)
KNN13	0.2688	0.2513	0.2239	0.2863	(**0.3137**, 130)
KNN15	0.2367	0.2075	0.1789	0.2317	(0.2505, 150)
KNN3	0.1843	0.1404	0.2933	0.2584	(0.2479, 160)
KNN5	0.2295	0.1475	**0.3241**	0.2746	(0.2392, 60)
KNN7	0.2317	0.2004	0.2619	0.2483	(**0.3293**, 400)
KNN9	0.2367	0.2641	0.2434	0.2619	(0.2118, 790)
SVM-C1	0.1901	0.2082	0.1952	0.1710	(0.1523, 50)
SVM-C10	0.1901	0.2082	0.2312	0.2250	(0.1708, 50)
SVM-C100	0.1901	0.2082	0.1952	0.2198	(0.1584, 50)

Columns are the same as in Table 3. MAQC teams best performance $MCC = 0.293$ and median 0.193. Performances superior to MAQC teams are highlighted in bold.

TABLE 7
Multiple Myeloma Dataset Divided According to a Random Outcome (CPR1)

	Suvrel	tt-all	tt-10	tt-100	loo
KNN11	0.0897	-0.0161	0.0716	0.0775	(0.0891, 50)
KNN13	0.0077	0.0302	0.0467	0.1166	(0.0758, 40)
KNN15	0.0677	0.0876	0.0276	0.0549	(0.0677, 40)
KNN3	-0.0068	-0.1008	0.0677	-0.0449	(0.0220, 130)
KNN5	-0.0545	-0.0765	0.0667	0.0020	(-0.0262, 70)
KNN7	-0.0424	-0.0575	0.0005	0.0188	(0.0465, 140)
KNN9	-0.0484	-0.0136	0.0194	0.0610	(0.1260, 130)
SVM-C1	-0.1138	-0.1015	-0.0025	-0.0065	(-0.0138, 30)
SVM-C10	-0.1138	-0.1015	0.0083	-0.0373	(-0.0141, 30)
SVM-C100	-0.1138	-0.1015	0.0110	-0.1228	(-0.0059, 30)

Columns are the same as in Table 3. In this case, since the classes are chosen randomly it is expected a MCC close to zero.

TABLE 8
Performance, Measured by MCC, for the Leukemia Dataset

	Suvrel	all	top10	top100
KNN3	0.9168	0.8390	0.8349	0.7958
KNN5	0.8769	0.8390	0.8349	0.8390
KNN7	0.8769	0.8390	0.8349	0.8390
KNN9	0.8769	0.7108	0.8349	0.8390
KNN11	0.8769	0.7224	0.7958	0.8390
KNN13	0.8769	0.7224	0.7958	0.8390
KNN15	0.8769	0.7224	0.7958	0.8390
SVM-C1	**0.9589**	0.9168	0.8349	0.8761
SVM-C10	**0.9589**	0.9168	0.8349	0.8761
SVM-C100	**0.9589**	0.9168	0.8349	0.8761

The best result is highlighted in bold.

ACKNOWLEDGMENTS

The authors thank Edgar Z. Alvarenga and Alexandre Patriota for discussions. This research was supported by the Center for the Study of Natural and Artificial Information Processing Systems of the University of São Paulo (CNAIPS, Nucleo de Apoio a Pesquisa da Universidade de Sao Paulo). MB and JEC received support from a FAPESP fellowship, VBPL had support from FAPESP and CNPq and NC thanks the IFT Foundation.

REFERENCES

[1] S. Leming, G. Campbell, W. D. Jones, F. Campagne, Z. Wen, S. J. Walker, Z. Su, T. M. Chu, F. M. Goodsaid, L. Pusztai, J. D. Jr, Shaughnessy, A. Oberthuer, R. S. Thomas, R. S. Paules, M. Fielden, B. Barlogie, W. Chen, P. Du, M. Fischer, C. Furlanello, B. D. Gallas, X. Ge, D. B. Megherbi, W. F. Symmans, M. D. Wang, J. Zhang, H. Bitter, B. Brors, P. R. Bushel, M. Bylesjo, M. Chen, J. Cheng, J. Cheng, J. Chou, T. S. Davison, M. Delorenzi, Y. Deng, V. Devanarayan, D. J. Dix, J. Dopazo, K. C. Dorff, F. Elloumi, J. Fan, S. Fan, X. Fan, H. Fang, N. Gonzaludo, K. R. Hess, H. Hong, J. Huan, R. A. Irizarry, R. Judson, D. Juraeva, S. Lababidi, C. G. Lambert, L. Li, Y. Li, Z. Li, S. M. Lin, G. Liu, E. K. Lobenhofer, J. Luo, W. Luo, M. N. McCall, Y. Nikolsky, G. A. Pennello, R. G. Perkins, R. Philip, V. Popovici, N. D. Price, F. Qian, A. Scherer, T. Shi, W. Shi, J. Sung, D. Thierry-Mieg, J. Thierry-Mieg, V. Thodima, J. Trygg, L. Vishnuvajjala, S. J. Wang, J. Wu, Y. Wu, Q. Xie, W. A. Yousef, L. Zhang, X. Zhang, S. Zhong, Y. Zhou, S. Zhu, D. Arasappan, W. Bao, A. B. Lucas, F. Berthold, R. J. Brennan, A. Buness, J. G. Catalano, C. Chang, R. Chen, Y. Cheng, J. Cui, W. Czika, F. Demichelis, X. Deng, D. Dosymbekov, R. Eils, Y. Feng, J. Fostel, S. Fulmer-Smentek, J. C. Fuscoe, L. Gatto, W. Ge, D. R. Goldstein, L. Guo, D. N. Halbert, J. Han, S. C. Harris, C. Hatzis, D. Herman, J. Huang, R. V. Jensen, R. Jiang, C. D. Johnson, G. Jurman, Y. Kahlert, S. A. Khuder, M. Kohl, J. Li, L. Li, M. Li, Q. Z. Li, S. Li, Z. Li, J. Liu, Y. Liu, Z. Liu, L. Meng, M. Madera, F. Martinez-Murillo, I. Medina, J. Meehan, K. Miclaus, R. A. Moffitt, D. Montaner, P. Mukherjee, G. J. Mulligan, P. Neville, T. Nikolskaya, B. Ning, G. P. Page, J. Parker, R. M. Parry, X. Peng, R. L. Peterson, J. H. Phan, B. Quanz, Y. Ren, S. Riccadonna, A. H. Roter, F. W. Samuelson, M. M. Schumacher, J. D. Shambaugh, Q. Shi, R. Shippy, S. Si, A. Smalter, C. Sotiriou, M. Soukup, F. Staedtler, G. Steiner, T. H. Stokes, Q. Sun, P. Y. Tan, R. Tang, Z. Tezak, B. Thorn, M. Tsyganova, Y. Turpaz, S. C. Vega, R. Visintainer, J. von Frese, C. Wang, E. Wang, J. Wang, W. Wang, F. Westermann, J. C. Willey, M. Woods, S. Wu, N. Xiao, J. Xu, L. Xu, L. Yang, X. Zeng, J. Zhang, L. Zhang, M. Zhang, C. Zhao, R. K. Puri, U. Scherf, W. Tong, R. D. Wolfinger, and MAQC Consortium, "The MicroArray Quality Control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models," *Nat. Biotechnol.*, vol. 28, no. 8, pp. 827–838, 2010.

[2] A.-C. Haury, P. Gestraud, and J.-P. Vert, "The influence of feature selection methods on accuracy, stability and interpretability of molecular signatures," *PloS One*, vol. 6, no. 12, p. e28210, 2011.

[3] B. Hammer and T. Villmann, "Generalized relevance learning vector quantization," *Neural Netw.*, vol. 15, nos. 8/9, pp. 1059–1068, 2002.

[4] P. Schneider, M. Biehl, and B. Hammer, "Adaptive relevance matrices in learning vector quantization," *Neural Comput.*, vol. 21, no. 12, pp. 3532–3561, 2009.

[5] W. Arlt, M. Biehl, A. E. Taylor, S. Hahner, R. Libe, B. A. Hughes, P. Schneider, D. J. Smith, H. Stiekema, N. Krone, E. Porfiri, G. Opocher, J. Bertherat, F. Mantero, B. Allolio, M. Terzolo, P. Nightingale, C. H. L. Shackleton, X. Bertagna, M. Fassnacht, and P. M. Stewart, "Urine steroid metabolomics as a biomarker tool for detecting malignancy in adrenal tumors," *J. Clin. Endocrinol. Metab.*, vol. 96, pp. 3775–3784, 2011.

[6] M. Kästner, B. Hammer, M. Biehl, and T. Villmann, "Functional relevance learning in generalized learning vector quantization," *Neurocomputing*, vol. 90, pp. 85–95, 2012.

[7] P. Schneider, K. Bunte, H. Stiekema, B. Hammer, T. Villmann, and M. Biehl, "Regularization in matrix relevance learning," *IEEE Trans. Neural Netw.*, vol. 21, no. 5, pp. 831–840, May 2010.

[8] K. Q. Weinberger, and L. K. Saul, "Distance metric learning for large margin," *J. Mach. Learn. Res.*, vol. 10, pp. 207–244, 2009.

[9] B. Kulis, "Metric learning: A survey," *Found. Trends Mach. Learn.*, vol. 5, no. 4, pp. 287–364, 2012.

[10] A. Bellet, A. Habrard, and M. Sebban, "A survey on metric learning for feature vectors and structured data," *CoRR*, vol. abs/1306.6709, 2013.

[11] M. Boareto, M. E. B. Yamagishi, N. Caticha, and V. B. P. Leite, "Relationship between global structural parameters and enzyme commission hierarchy: Implications for function prediction," *Comput. Biol. Chem.*, vol. 40, pp. 15–19, 2012.

[12] P. C. Mahalanobis, "On the generalized distance in statistics," *Proc. Nat. Inst. Sci. India*, vol. 2, no. 1, p. 4955, 1936.

[13] J. Gorodkin, "Comparing two k-category assignments by a k-category correlation coefficient," *Comput. Biol. Chem.*, vol. 28, nos. 5/6, pp. 367–374, 2004.

[14] G. Jurman, S. Riccadonna, and C. Furlanello, "A comparison of mcc and cen error measures in multi-class prediction," *PloS One*, vol. 7, no. 8, p. e41882, 2012.

[15] R. A. Irizarry, B. Hobbs, F. Collin, Y. D. Beazer-Barclay, K. J. Antonellis, U. Scherf, and T. P. Speed, "Exploration, normalization, and summaries of high density oligonucleotide array probe level data," *Biostatistics*, vol. 4, no. 2, pp. 249–264, 2003.

[16] S. A. Armstrong, J. E. Staunton, L. B. Silverman, R. Pieters, M. L. den Boer, M. D. Minden, S. E. Sallan, E. S. Lander, T. R. Golub, and S. J. Korsmeyer, "MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia," *Nat. Genet.*, vol. 30, no. 1, pp. 41–47, 2002.

[17] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.

[18] B. N. Ametaj, Q. Zebeli, F. Saleem, N. Psychogios, M. J. Lewis, S. M. Dunn, J. Xia, and D. S. Wishart, "Metabolomics reveals unhealthy alterations in rumen metabolism with increased proportion of cereal grain in the diet of dairy cows," *Metabolomics*, vol. 6, no. 4, pp. 583–594, 2010.

[19] L. Ein-Dor, O. Zuk, and E. Domany, "Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer," *Proc. Nat. Acad. Sci. USA*, vol. 103, no. 15, pp. 5923–5928, 2006.

[20] L. Ein-Dor, I. Kela, G. Getz, D. Givol, and E. Domany, "Outcome signature genes in breast cancer: Is there a unique set?" *Bioinformatics*, vol. 21, no. 2, pp. 171–178, 2005.