Multimodal Omics Data Integration Using Max Relevance–Max Significance Criterion

Pradipta Maji*, *Senior Member, IEEE*, and Ankita Mandal

Abstract*—*Objective: **This paper presents a novel supervised regularized canonical correlation analysis, termed as CuRSaR, to extract relevant and significant features from multimodal high dimensional omics datasets.** ***Methods:*** **The proposed method extracts a new set of features from two multidimensional datasets by maximizing the relevance of extracted features with respect to sample categories and significance among them. It integrates judiciously the merits of regularized canonical correlation analysis (RCCA) and rough hypercuboid approach. An analytical formulation, based on spectral decomposition, is introduced to establish the relation between canonical correlation analysis (CCA) and RCCA. The concept of hypercuboid equivalence partition matrix of rough hypercuboid is used to compute both relevance and significance of a feature.** ***Significance:*** **The analytical formulation makes the computational complexity of the proposed algorithm significantly lower than existing methods. The equivalence partition matrix offers an efficient way to find optimum regularization parameters employed in CCA.** ***Results:*** **The superiority of the proposed algorithm over other existing methods, in terms of computational complexity and classification accuracy, is established extensively on real life data.**

***Index Terms*—Canonical correlation analysis (CCA), classification, feature extraction, multimodal data analysis, rough sets.**

I. Introduction

INTEGRATION of different omics data, such as gene expression array, protein expression array, microRNA array, methylation, and copy number variation data, may provide a better understanding of the biological systems. The simultaneous analysis of such transcriptions, proteomics or metabolomics is an important task in integrative systems biology approach. It gives better understanding of the relationships among different biological functional levels. Due to the drastic variation and noisy nature of the acquired signals, unimodal-based pattern analysis and recognition systems usually afford low level of performance, which lead to insufficient and inaccurate pattern representation of the perception of interest. On the other hand, multimodal data contain more information. By using multiple types of data of unique sample, it is possible to make the linkages between attributes within each type of data. The combination of multimodal data may potentially provide a more complete and discriminatory description of the intrinsic characteristics of the pattern by producing improved system performance than single modality only.

Canonical correlation analysis (CCA) [1] is a bivariate feature extraction method, which provides an efficient way of measuring the linear relationship between two multidimensional variables. The goal of CCA is to find the best linear transformation for two multidimensional datasets so that the maximum correlation between them can be achieved. Recently, there has been an increasing interest in applying CCA to many important fields of biomedical sciences [2]–[7]. To integrate different omics data, CCA has been widely applied [8], [9]. Several variants of the CCA method, such as kernel CCA [2], [10] and sparse CCA [11], [12] have also been developed in recent years. In [2], kernel CCA has been used to map the genes or proteins onto the Euclidean space, where connected nodes are close to each other, while Yamanishi *et al.* [10] used kernel CCA to infer enzyme networks from the integration of multiple genomic data and chemical information. On the other hand, sparse CCA is used to study the mutual relationship between two different types of genomic data [11], [12]. Besides of integrating two datasets, CCA can help to analyze gene expression dynamics geometrically [13]. Phylogenetic CCA [14], another variant of CCA, gives continuous valued character data obtained from biological species related by a phylogenetic tree. Hence, CCA can be used to capture the underlying genetic background of a complex disease, by associating two datasets containing information about a patient's phenotypical and genetic details. It gives those relevant variables or features from both data types, which are related to each other and provide more insight into the biological experimental hypotheses.

Let $\mathcal{X}$ and $\mathcal{Y}$ be two multivariate datasets having p and q number of features, respectively, and n is the number of samples in both $\mathcal{X}$ and $\mathcal{Y}$. In bioinformatics domain, the number of training samples n is usually limited, while the modern technology has enabled very high-dimensional data streams to be routinely acquired, which results in very high-dimensional feature spaces p and q. When $n << p$ and $n << q$, the features in $\mathcal{X}$ and $\mathcal{Y}$ tend to be highly collinear, which leads to ill-conditioned of the covariance matrices C_{xx} and C_{yy} of $\mathcal{X}$ and $\mathcal{Y}$, respectively. In effect, their inverses are no longer reliable, resulting in an invalid computation of CCA and an unreliable meta-space [15]. To overcome this problem, a regularized version of CCA

Manuscript received September 8, 2016; revised October 20, 2016; accepted October 28, 2016. Date of publication November 4, 2016; date of current version July 15, 2017. This work was supported in part by the Department of Electronics and Information Technology, Government of India (PhD-MLA/4(90)/2015-16). *(Asterisk indicates corresponding author.)*

*P. Maji is with the Biomedical Imaging and Bioinformatics Lab, Machine Intelligence Unit, Indian Statistical Institute, Kolkata 700108, India (e-mail: pmaji@isical.ac.in).

A. Mandal is with the Biomedical Imaging and Bioinformatics Lab, Machine Intelligence Unit, Indian Statistical Institute.

Digital Object Identifier 10.1109/TBME.2016.2624823

is introduced in [16]. Regularized CCA (RCCA) [17], [18] is an improved version of CCA. It prevents overfitting of insufficient training data by using a ridge regression optimization scheme [19]. It works by adding small positive quantities to the diagonals of C_{xx} and C_{yy} to guarantee their invertibility [20]. In [21], an alternative method to the existing RCCA has been presented, which is based on the estimates of the correlation matrices that minimize the mean squared error risk function. An *et al*. [22] proposed a robust CCA, which uses shrinkage estimation and smoothing technique to estimate the data covariance matrices with limited samples. RCCA has been successfully used to study gene expressions in liver cells and compare them with concentrations of hepatic fatty acids in mice [18]. Regularized sparse CCA is used in expression quantitative trait loci to detect genetic loci mapped to a disease [23]. However, RCCA is computationally very expensive because of this regularization process. Also, both CCA and RCCA are unsupervised and fail to take complete advantage of available class label information [24].

Supervised RCCA (SRCCA) incorporates a supervised feature selection scheme to perform the regularization [24], [25]. It includes the information of available class label to select maximally correlated features. In SRCCA, regularization is done by embedding component with the most discriminatory score as chosen by feature selection scheme and then adjusted for the remaining dimensions [24], [25]. It only considers the correlation of first pair of canonical variables. But, it may happen that other canonical variable pairs have insignificant relation with first pair of canonical variables, or there may be some irrelevant features in the whole extracted feature set, which should not be considered in further processing [26]. Moreover, uncertainty in omics data analysis is one of the major concerns. Some of the sources of this uncertainty include imprecision in computation and vagueness in class definition. The t-test, Wilcoxon rank sum test or Wilks's lambda test, used to capture supervised class information in SRCCA [24], are unable to handle this uncertainty. In this context, the rough set theory has gained popularity in modeling and propagating uncertainty. It deals with vagueness and incompleteness, and is proposed for indiscernibility in classification, according to some similarity. It has been applied successfully to feature selection and clustering [27] as well as to omics data analysis [26]–[30].

In this regard, the paper presents a new feature extraction algorithm, termed as CuRSaR (CCA using maximum Relevance-maximum Significance criterion and Rough sets), from two multidimensional datasets. It judiciously integrates the merits of SRCCA and the theory of rough sets. The proposed method extracts a set of new features by maximizing their relevance with respect to class labels and significance among them [28]. Both the relevance and significance measures are computed based on the concept of hypercuboid equivalence partition matrix of rough hypercuboid approach [29]. In the proposed method, the regularization parameters do not only depend on the first pair of canonical variables, rather the whole extracted feature set is considered to optimize the regularization parameters. An analytical formulation is presented to establish the relation between CCA and RCCA, which makes the computational cost of the proposed algorithm significantly lower than existing methods. The effectiveness of the proposed method, along with a comparison with other methods, is demonstrated on several real life datasets.

II. Basics of CCA

CCA [1] obtains a linear relationship between two multidimensional variables. The objective of CCA is to extract latent features from two datasets $\mathcal{X} \in \mathbb{R}^{p\times n}$ and $\mathcal{Y} \in \mathbb{R}^{q\times n}$, which are highly correlated, where each column in $\mathcal{X}$ and $\mathcal{Y}$ corresponds to one of the n samples, and each row represents one variable. Let us assume that each variable is centered to have zero mean across the samples. CCA obtains two-directional weight vectors, also termed as basis vectors, $w_x \in \mathbb{R}^p$ and $w_y \in \mathbb{R}^q$ such that the empirical correlation between the respective projections onto these weight vectors, that is, between $\mathcal{X}^T w_x$ and $\mathcal{Y}^T w_y$ is maximum. The correlation coefficient $\tilde{\rho}$ is given as follows:

$$\tilde{\rho} = \max_{w_x, w_y} \frac{w_x^T C_{xy} w_y}{\sqrt{w_x^T C_{xx} w_x w_y^T C_{yy} w_y}} \tag{1}$$

where $C_{xy} \in \mathbb{R}^{p\times q}$ is the cross-covariance matrix of $\mathcal{X}$ and $\mathcal{Y}$, while $C_{xx} \in \mathbb{R}^{p\times p}$ and $C_{yy} \in \mathbb{R}^{q\times q}$ are covariance matrices of $\mathcal{X}$ and $\mathcal{Y}$, respectively. Since $\tilde{\rho}$ is invariant to the scaling of w_x and w_y, CCA can be formulated equivalently as

$$\max_{w_x, w_y} \quad w_x^T C_{xy} w_y$$
$$\text{subject to} \quad w_x^T C_{xx} w_x = 1; \quad w_y^T C_{yy} w_y = 1. \tag{2}$$

To calculate w_x and w_y, eigenvectors of $\Sigma\Sigma^T$ and $\Sigma^T\Sigma$ are needed, where matrix $\Sigma \in \mathbb{R}^{p\times q}$ is given as follows:

$$\Sigma = C_{xx}^{-1/2} C_{xy} C_{yy}^{-1/2}. \tag{3}$$

If ξ_{xi} and ξ_{yi}, respectively, are the orthonormalized eigenvectors of $\Sigma\Sigma^T$ and $\Sigma^T\Sigma$ corresponding to ith eigenvalue ρ_i, then the ith pair of basis vectors are given by

$$w_{xi} = C_{xx}^{-1/2} \xi_{xi}; \quad \text{and} \quad w_{yi} = C_{yy}^{-1/2} \xi_{yi} \tag{4}$$

and the ith pair of canonical variables are as follows:

$$\mathcal{U}_i = w_{xi}^T \mathcal{X}; \quad \text{and} \quad \mathcal{V}_i = w_{yi}^T \mathcal{Y}. \tag{5}$$

Here, $(\mathcal{U}_1, \mathcal{V}_1)$ is the first pair of canonical variables, which provides the maximum correlation $\tilde{\rho} = \sqrt{\rho_1}$. The ith pair of canonical variables $(\mathcal{U}_i, \mathcal{V}_i)$ is the linear combinations of dataset and basis vector having unit variance. It maximizes the correlation among all possible linear combinations and is uncorrelated with the previous $(i\text{-}1)$ canonical variable pairs. From (5), the ith feature $\mathcal{A}_i$ is extracted as follows:

$$\mathcal{A}_i = \mathcal{U}_i + \mathcal{V}_i \tag{6}$$

where $\forall i \in \{1, 2, \ldots, \mathcal{K}\}$ and $\mathcal{K} \leqslant \min(p, q)$.

When $n << p$ and $n << q$, regularized CCA (RCCA) [17], [18] adds small positive quantities to the diagonals of C_{xx} and C_{yy} to guarantee their invertibility [20]. So, (3) becomes

$$\Sigma = [C_{xx} + \mathfrak{r}_x I]^{-1/2} C_{xy} [C_{yy} + \mathfrak{r}_y I]^{-1/2} \tag{7}$$

where I is the identity matrix, $\mathfrak{r}_x$ and $\mathfrak{r}_y$ are known as regularization parameters, which are varied in a certain range $\mathfrak{r}_{min} \leqslant \mathfrak{r}_x, \mathfrak{r}_y \leqslant \mathfrak{r}_{max}$ and chosen by a grid search optimization technique [31]. Every pair of $\mathfrak{r}_x$ and $\mathfrak{r}_y$ will produce a pair of first canonical variables, which are maximally correlated. The optimal parameters $\mathfrak{r}_x$ and $\mathfrak{r}_y$ are considered for which the Pearson's

correlation is maximum, that is,

$$\max_{\mathfrak{r}_x,\mathfrak{r}_y} \frac{w_x^T C_{xy} w_y}{\sqrt{w_x^T [C_{xx}+\mathfrak{r}_x I] w_x w_y^T [C_{yy}+\mathfrak{r}_y I] w_y}}. \tag{8}$$

III. Proposed Method

This section presents a new feature extraction algorithm, termed as CuRSaR, integrating judiciously the information of two multidimensional datasets. Prior to describing the proposed method for multimodal data analysis, some important analytical formulations are introduced next, which reduce the computational complexity of existing RCCA.

A. Regularized CCA

To calculate $[C_{xx}+\mathfrak{r}_x I]^{-1/2}$ and $[C_{yy}+\mathfrak{r}_y I]^{-1/2}$ for the computation of Σ matrix of (7), spectral decomposition can be used [32]. The spectral decomposition can be described in terms of eigenvalue-eigenvector pairs of $[C_{xx}+\mathfrak{r}_x I]$ and $[C_{yy}+\mathfrak{r}_y I]$. A $p \times p$ symmetric matrix $[C_{xx}+\mathfrak{r}_x I]$ can be expressed in terms of its p eigenvalue-eigenvector pairs (Λ_x, Ψ_x) as follows [32]:

$$[C_{xx}+\mathfrak{r}_x I] = \Psi_x \Lambda_x \Psi_x^T = \sum_{i=1}^{p} \lambda_i \psi_i \psi_i^T \tag{9}$$

where λ_i denotes the diagonal elements of Λ_x and ψ_i be the columns of matrix Ψ_x, which are the orthonormalized eigenvectors of eigenvalues $\lambda_i, \forall i \in \{1,2,\ldots,p\}$, and

$$\Psi_x \Psi_x^T = \Psi_x^T \Psi_x = I. \tag{10}$$

The computation of the inverse square root matrix is performed as follows [33]:

$$[C_{xx}+\mathfrak{r}_x I]^{-1/2} = \Psi_x \Lambda_x^{-1/2} \Psi_x^T = \sum_{i=1}^{p} \frac{1}{\sqrt{\lambda_i}} \psi_i \psi_i^T. \tag{11}$$

In RCCA and SRCCA, the regularization parameters $\mathfrak{r}_x$ and $\mathfrak{r}_y$ are varied within a specified range $[\mathfrak{r}_{min}, \mathfrak{r}_{max}]$, where $\mathfrak{r}_{min} \leqslant \mathfrak{r}_x, \mathfrak{r}_y \leqslant \mathfrak{r}_{max}$. It can be assumed that these regularization parameters follow an arithmetic progression. Each parameter starts with an initial value $\mathfrak{r}_{min}$. After every iteration, a constant value or a common difference is added with the previous value, and finally, it reaches to $\mathfrak{r}_{max}$. Let us assume that d_x and d_y be the common differences for $\mathfrak{r}_x$ and $\mathfrak{r}_y$, respectively. So, the arithmetic progression series can be thought as follows:

$$\begin{aligned} &\mathfrak{r}_x, \mathfrak{r}_x + d_x, \ldots, \mathfrak{r}_x + i d_x, \ldots, \mathfrak{r}_x + (\mathfrak{t}_x - 1) d_x \\ &\mathfrak{r}_y, \mathfrak{r}_y + d_y, \ldots, \mathfrak{r}_y + j d_y, \ldots, \mathfrak{r}_y + (\mathfrak{t}_y - 1) d_y \end{aligned} \tag{12}$$

where initially $\mathfrak{r}_x = \mathfrak{r}_{min}$ and $\mathfrak{r}_y = \mathfrak{r}_{min}$ and at final step $\mathfrak{r}_x + (\mathfrak{t}_x - 1)d_x = \mathfrak{r}_{max}$ and $\mathfrak{r}_y + (\mathfrak{t}_y - 1)d_y = \mathfrak{r}_{max}$. The parameters $\mathfrak{t}_x$ and $\mathfrak{t}_y$ denote the number of possible values of regularization parameters $\mathfrak{r}_x$ and $\mathfrak{r}_y$, respectively. It is clearly seen that the diagonal elements of the covariance matrices are only changed by adding regularization parameters. Let $[C_{xx}+\mathfrak{r}_x I]$ has eigenvalue λ_{x_i} and eigenvector ψ_x. So

$$[C_{xx}+\mathfrak{r}_x I]\psi_x = \lambda_{x_i} \psi_x. \tag{13}$$

Let us assume that a scalar d_x is added on the diagonal elements of the matrix $[C_{xx}+\mathfrak{r}_x I]$. Multiplying this new matrix by the vector ψ_x, we get

$$\begin{aligned} [C_{xx}+(\mathfrak{r}_x + d_x)I]\psi_x &= [C_{xx}+\mathfrak{r}_x I]\psi_x + d_x I \psi_x \\ &= \lambda_{x_i}\psi_x + d_x \psi_x = (\lambda_{x_i} + d_x)\psi_x. \end{aligned} \tag{14}$$

Hence, if a regularization parameter is added on the diagonal elements of the covariance matrix, the eigenvalues are changed, but the eigenvectors remain same.

Let $\Lambda_{x_1}, \ldots, \Lambda_{x_{i+1}}, \ldots, \Lambda_{x_{\mathfrak{t}_x}}$ be the diagonal matrices, where diagonal elements are the eigenvalues of $[C_{xx}+\mathfrak{r}_x I], \ldots, [C_{xx}+(\mathfrak{r}_x + i d_x)I], \ldots, [C_{xx}+(\mathfrak{r}_x+(\mathfrak{t}_x-1)d_x)I]$. Similarly, $\Lambda_{y_1}, \ldots, \Lambda_{y_{j+1}}, \ldots, \Lambda_{y_{\mathfrak{t}_y}}$ are the diagonal matrices with eigenvalues of $[C_{yy}+\mathfrak{r}_y I], \ldots, [C_{yy}+(\mathfrak{r}_y + j d_y)I], \ldots, [C_{yy}+(\mathfrak{r}_y+(\mathfrak{t}_y-1)d_y)I]$ on the diagonal elements. The corresponding orthonormal eigenvectors are in the columns of Ψ_x and Ψ_y. So, eigenvalue-eigenvector equations can be written as follows:

$$[C_{xx}+(\mathfrak{r}_x + (i-1)d_x)I]\Psi_x = \Psi_x \Lambda_{x_i} \tag{15}$$

$$[C_{yy}+(\mathfrak{r}_y + (j-1)d_y)I]\Psi_y = \Psi_y \Lambda_{y_j} \tag{16}$$

where $\forall i \in \{1,2,\ldots,\mathfrak{t}_x\}$ and $\forall j \in \{1,2,\ldots,\mathfrak{t}_y\}$. From (15), we get

$$\begin{aligned} &[C_{xx}+\mathfrak{r}_x I]\Psi_x + (i-1)d_x I \Psi_x = \Psi_x \Lambda_{x_i} \\ &\Rightarrow \Psi_x \Lambda_{x_1} + (i-1)d_x \Psi_x = \Psi_x \Lambda_{x_i} \\ &\Rightarrow \Psi_x(\Lambda_{x_i} - \Lambda_{x_1} - (i-1)d_x I) = 0 \\ &\Rightarrow \Lambda_{x_i} = \Lambda_x + (i-1)d_x I; \end{aligned} \tag{17}$$

where $\Lambda_x = \Lambda_{x_1}$. Similarly, from (16), we get

$$\Lambda_{y_j} = \Lambda_y + (j-1)d_y I; \tag{18}$$

where $\Lambda_y = \Lambda_{y_1}$. Combining (15), (17) and (16), (18), we get

$$[C_{xx}+(\mathfrak{r}_x + i d_x)I]\Psi_x = \Psi_x(\Lambda_x + i d_x I); \tag{19}$$

$$[C_{yy}+(\mathfrak{r}_y + j d_y)I]\Psi_y = \Psi_y(\Lambda_y + j d_y I). \tag{20}$$

From (19) and (20), it is clearly seen that there is no need to calculate eigenvalue for every pair of regularization parameters $\mathfrak{r}_x$ and $\mathfrak{r}_y$. It is sufficient to calculate eigenvalues Λ_x and Λ_y for the initial values of $\mathfrak{r}_x$ and $\mathfrak{r}_y$, respectively. The eigenvalues corresponding to other values of $\mathfrak{r}_x$ and $\mathfrak{r}_y$ can be computed from the initial values using (17) and (18). On the other hand, relations (14), (19), and (20) establish the fact that eigenvectors remain unchanged irrespective of the values of regularization parameters. Moreover, if the minimum value of regularization parameter $\mathfrak{r}_{min}$ is set to 0, then RCCA with regularization parameters $\mathfrak{r}_x = \mathfrak{r}_y = 0$ reduces to classical CCA. So, for $\mathfrak{r}_{min} = 0$, the eigenvalues and eigenvectors of CCA can be used to compute eigenvalues and eigenvectors of RCCA, corresponding to other values $\mathfrak{r}_x$ and $\mathfrak{r}_y$, using (17)–(20).

Also, if the regularization parameters $\mathfrak{r}_x$ and $\mathfrak{r}_y$ follow an arithmetic progression as in (12), the matrix Σ of (7) becomes

$$\begin{aligned} \Sigma_{ij} = \; &[C_{xx}+(\mathfrak{r}_x + (i-1)d_x)I]^{-1/2} C_{xy} \\ &[C_{yy}+(\mathfrak{r}_y + (j-1)d_y)I]^{-1/2} \end{aligned} \tag{21}$$

where $\forall i \in \{1,2,\ldots,\mathfrak{t}_x\}$ and $\forall j \in \{1,2,\ldots,\mathfrak{t}_y\}$. Combining (11), (19), (20), and (21), we get

$$\begin{aligned} \Sigma_{ij} = \; &\Psi_x(\Lambda_x + (i-1)d_x I)^{-1/2} \Psi_x^T C_{xy} \\ &\Psi_y(\Lambda_y + (j-1)d_y I)^{-1/2} \Psi_y^T. \end{aligned} \tag{22}$$

From (22), it is clear that if eigenvalues and eigenvectors are calculated to compute Σ_{11} matrix for initial values of $\mathfrak{r}_x$ and $\mathfrak{r}_y$, there is no need to compute eigenvalues and eigenvectors for computing Σ_{ij} at other values of $\mathfrak{r}_x$ and $\mathfrak{r}_y$, where $\forall i \in \{1,2,\ldots,\mathfrak{t}_x\}$ and $\forall j \in \{1,2,\ldots,\mathfrak{t}_y\}$, as initial eigenvalues and eigenvectors can be used to compute different Σ_{ij} matrix. Also, if the minimum value of $\mathfrak{r}_x$ and $\mathfrak{r}_y$ is set to 0, then eigenvalues and eigenvectors of CCA can be used to compute different Σ_{ij} matrix of RCCA corresponding to different values of regularization parameters.

After computing Σ matrix corresponding to a pair of regularization parameters $\mathfrak{r}_x$ and $\mathfrak{r}_y$ using (22), either $\Sigma\Sigma^T$ or $\Sigma^T\Sigma$ is computed depending on whether $p \leqslant q$ or $p > q$. As nonzero eigenvalues of $\Sigma\Sigma^T$ are same as nonzero eigenvalues of $\Sigma^T\Sigma$ [34], one of the matrices is enough to calculate the eigenvector of $\Sigma\Sigma^T$ or $\Sigma^T\Sigma$, where

$$\Sigma\Sigma^T = C_{xx}^{-1/2} C_{xy} C_{yy}^{-1} C_{yx} C_{xx}^{-1/2} \tag{23}$$

$$\text{and} \quad \Sigma^T\Sigma = C_{yy}^{-1/2} C_{yx} C_{xx}^{-1} C_{xy} C_{yy}^{-1/2}. \tag{24}$$

Suppose $\rho_1 \geqslant \cdots \geqslant \rho_k \geqslant \cdots \geqslant \rho_p$ are the eigenvalues of $\Sigma\Sigma^T$ and $\xi_{x1},\ldots,\xi_{xk},\ldots,\xi_{xp}$ are the orthonormalized eigenvectors corresponding to $\rho_1,\ldots,\rho_k,\ldots,\rho_p$. Furthermore, let say, $p < q$ and $\rho_1 \geqslant \cdots \geqslant \rho_k \geqslant \cdots \geqslant \rho_p$ are the p largest eigenvalues of $\Sigma^T\Sigma$ with orthonormalized eigenvectors $\xi_{y1},\ldots,\xi_{yk},\ldots,\xi_{yp}$. So

$$\begin{aligned} \Sigma\Sigma^T \xi_{xk} = \rho_k \xi_{xk}; \;&\Rightarrow\; \Sigma^T\Sigma\Sigma^T \xi_{xk} = \rho_k \Sigma^T \xi_{xk}; \\ &\Rightarrow\; \Sigma^T\Sigma \xi_{yk} = \rho_k \xi_{yk}. \end{aligned} \tag{25}$$

The kth eigenvector ξ_{yk} of $\Sigma^T\Sigma$ is proportional to $C_{yy}^{-1/2} C_{yx} C_{xx}^{-1/2} \xi_{xk}$, that is, $\xi_{yk} = \Sigma^T \xi_{xk}$. From (25), it can also be seen that either $\Sigma\Sigma^T$ or $\Sigma^T\Sigma$ is enough to calculate the eigenvector of $\Sigma\Sigma^T$ and $\Sigma^T\Sigma$. For RCCA with (i,j)th regularization parameters of $\mathfrak{r}_x$ and $\mathfrak{r}_y$, the $\Sigma_{ij}\Sigma_{ij}^T$ and $\Sigma_{ij}^T\Sigma_{ij}$ are calculated as follows:

$$\begin{aligned} \Sigma_{ij}\Sigma_{ij}^T = \;& \Psi_x(\Lambda_x + (i-1)d_x I)^{-1/2} \Psi_x^T C_{xy} \Psi_y (\Lambda_y \\ & + (j-1)d_y I)^{-1} \Psi_y^T C_{yx} \Psi_x (\Lambda_x + (i-1)d_x I)^{-1/2} \Psi_x^T; \end{aligned} \tag{26}$$

$$\begin{aligned} \Sigma_{ij}^T\Sigma_{ij} = \;& \Psi_y(\Lambda_y + (j-1)d_y I)^{-1/2} \Psi_y^T C_{yx} \Psi_x (\Lambda_x \\ & + (i-1)d_x I)^{-1} \Psi_x^T C_{xy} \Psi_y (\Lambda_y + (j-1)d_y I)^{-1/2} \Psi_y^T. \end{aligned} \tag{27}$$

Assuming $\mathcal{K} = \min(p,q)$, $\mathcal{K}$ eigenvalue-eigenvector pairs can be calculated using Jacobi method [35]. Then, $\mathcal{K}$ pairs of basis vectors and $\mathcal{K}$ pairs of canonical variables are computed using (4) and (5), respectively. Finally, $\mathcal{K}$ features can be extracted using (6). The computational complexity of Jacobi method to compute $\mathcal{K}$ eigenvalue-eigenvector pairs is $\mathcal{O}(\mathcal{K}^3)$.

B. CuRSaR: Proposed Algorithm

One of the main problems in real life high dimensional multimodal data analysis is how to extract relevant and significant features. In general, the extracted feature set may contain a huge number of irrelevant and insignificant features. The presence of such features may lead to a reduction in the useful information and degrade the prediction capability. Thus, the extracted feature subset should contain the features which have high relevance and high significance in the feature set. Such features are expected to be able to predict the classes of the samples. Accordingly, a measure is required that can assess the effectiveness of a feature set. In this work, hypercuboid equivalence partition matrix of rough hypercuboid approach [29] is used to select relevant and significant features, which are extracted from two multidimensional datasets by calculating their maximum correlation and variation.

Let $\mathcal{X} \in \mathbb{R}^{p\times n}$ and $\mathcal{Y} \in \mathbb{R}^{q\times n}$ be two multidimensional datasets with p and q variables or attributes, respectively, and n samples. Let us assume that each variable is centered to have zero mean across the samples. Let $\mathfrak{t}_x$ and $\mathfrak{t}_y$ be the number of possible values of regularization parameters $\mathfrak{r}_x$ and $\mathfrak{r}_y$, respectively. The value of each regularization parameter is varied within a certain range $[\mathfrak{r}_{min}, \mathfrak{r}_{max}]$ as per (12), where $\mathfrak{r}_{min} \leqslant \mathfrak{r}_x, \mathfrak{r}_y \leqslant \mathfrak{r}_{max}$. Let $\mathcal{A}_{k_{ij}}$ be the kth extracted feature with (i,j)th regularization parameters of $\mathfrak{r}_x$ and $\mathfrak{r}_y$ and $\gamma_{\mathcal{A}_k}(\mathbb{D})$ be the relevance of the feature $\mathcal{A}_k$ with respect to the class labels $\mathbb{D}$. Define $\sigma_{\{\mathcal{A}_k,\mathcal{A}_l\}}(\mathbb{D},\mathcal{A}_k)$ as the significance of the feature $\mathcal{A}_k$ with respect to another feature $\mathcal{A}_l \in \mathbb{S}$, where $\mathbb{S}$ is the set of $\mathcal{D}$ selected features and $\mathcal{D} \leqslant \min(p,q)$. The change in dependency when a feature is removed from the set of features, is a measure of the significance of the feature. To what extent a feature is contributing to calculate the dependency on class labels can be calculated by the significance of that feature. The significance of the feature $\mathcal{A}_k$ with respect to the feature set $\{\mathcal{A}_k,\mathcal{A}_l\}$ is given by

$$\sigma_{\{\mathcal{A}_k,\mathcal{A}_l\}}(\mathbb{D},\mathcal{A}_k) = \gamma_{\{\mathcal{A}_k,\mathcal{A}_l\}}(\mathbb{D}) - \gamma_{\mathcal{A}_l}(\mathbb{D}). \tag{28}$$

Hence, the higher the change in dependency, the more significant the feature $\mathcal{A}_k$ is. If significance is 0, then the feature is dispensable. Therefore, the total relevance of all selected features for (i,j)th regularization parameters of $\mathfrak{r}_x$ and $\mathfrak{r}_y$ is given by

$$R(i,j) = \sum_{\mathcal{A}_{k_{ij}} \in \mathbb{S}} \gamma_{\mathcal{A}_{k_{ij}}}(\mathbb{D}) \tag{29}$$

while the total significance among the selected features is as follows

$$S(i,j) = \sum_{\mathcal{A}_{k_{ij}} \neq \mathcal{A}_{l_{ij}} \in \mathbb{S}} \sigma_{\{\mathcal{A}_{k_{ij}},\mathcal{A}_{l_{ij}}\}}(\mathbb{D},\mathcal{A}_{k_{ij}}) + \sigma_{\{\mathcal{A}_{k_{ij}},\mathcal{A}_{l_{ij}}\}}(\mathbb{D},\mathcal{A}_{l_{ij}}). \tag{30}$$

Therefore, the problem of extracting a set $\mathbb{S}$ of relevant and significant features from all possible combinations of regularization parameters $\mathfrak{r}_x$ and $\mathfrak{r}_y$ is equivalent to maximize both $R(i,j)$ and $S(i,j)$, that is, to maximize the objective function $J(i,j)$, where

$$J(i,j) = \omega \times \frac{R(i,j)}{\mathcal{D}} + (1-\omega) \times \frac{S(i,j)}{\mathcal{D}(\mathcal{D}-1)} \tag{31}$$

where ω is a weight parameter. The criterion combining the above two constraints is called *maximum relevance-maximum significance* [28], [29]. To solve the above problem, the following algorithm is used:

1) Calculate the cross-covariance matrix C_{xy} of $\mathcal{X}$ and $\mathcal{Y}$.

2) Calculate covariance matrices C_{xx} and C_{yy} of $\mathcal{X}$ and $\mathcal{Y}$, respectively.
3) Calculate eigenvalues $\Lambda_x \in \mathbb{R}^p$ and $\Lambda_y \in \mathbb{R}^q$ of C_{xx} and C_{yy}, along with corresponding eigenvectors Ψ_x and Ψ_y using Jacobi method.
4) Initialize $\mathbb{S} \leftarrow \emptyset$ and $J_{\text{optimal}} = 0$.
5) Repeat the following seven steps for all (i,j)th regularization parameters of $\mathfrak{r}_x$ and $\mathfrak{r}_y$, where $\forall i \in \{1, 2, \ldots, \mathfrak{t}_x\}$ and $\forall j \in \{1, 2, \ldots, \mathfrak{t}_y\}$.
   a) If $p \leqslant q$ (respectively, $q > p$), calculate $\Sigma_{ij}\Sigma_{ij}^T$ using (26) [respectively, $\Sigma_{ij}^T\Sigma_{ij}$ using (27)].
   b) Calculate first $\mathcal{D}$ eigenvectors $\{\xi_{x_{ij}}\}$ (respectively, $\{\xi_{y_{ij}}\}$) of $\Sigma_{ij}\Sigma_{ij}^T$ (respectively, $\Sigma_{ij}^T\Sigma_{ij}$) using Power and Deflation methods.
   c) Calculate $\xi_{y_{ij}} = \Sigma_{ij}^T \xi_{x_{ij}}$ (respectively, $\xi_{x_{ij}} = \Sigma_{ij}\xi_{y_{ij}}$).
   d) Calculate $\mathcal{D}$ pairs of basis vectors $\{w_{x_{ij}}, w_{y_{ij}}\}$ and $\mathcal{D}$ pairs of canonical variables $\{\mathcal{U}_{ij}, \mathcal{V}_{ij}\}$ using (4) and (5), respectively.
   e) Extract $\mathcal{D}$ features $\{\mathcal{A}_{ij}\}$ corresponding to (i,j)th pair of regularization parameters using (6) and store them in $\mathbb{C}$.
   f) Compute the objective function $J(i, j)$ using (31).
   g) If $J(i, j) > J_{\text{optimal}}$, then $\mathbb{S} \leftarrow \mathbb{C}$, and $J_{\text{optimal}} = J(i, j)$.
6) Output $\mathbb{S}$.
7) Stop.

C. Computation of Relevance and Significance

The concept of hypercuboid equivalence partition matrix of rough hypercuboid approach [29] is used to compute both the relevance and significance of an extracted feature. The regularization parameters are optimized through computing the relevance and significance measures. The relevance $\gamma_{\mathcal{A}_k}(\mathbb{D})$ of a feature or condition attribute $\mathcal{A}_k$ with respect to the class labels or decision attribute $\mathbb{D}$ is computed as follows [29]:

$$\gamma_{\mathcal{A}_k}(\mathbb{D}) = 1 - \frac{1}{n}\sum_{j=1}^{n} v_j(\mathcal{A}_k) \tag{32}$$

where n is the number of samples and

$$\mathbb{V}(\mathcal{A}_k) = [v_1(\mathcal{A}_k), \ldots, v_j(\mathcal{A}_k), \ldots, v_n(\mathcal{A}_k)] \tag{33}$$

is termed as the confusion vector for the attribute $\mathcal{A}_k$, where

$$v_j(\mathcal{A}_k) = \min\left\{1, \sum_{i=1}^{c} h_{ij}(\mathcal{A}_k) - 1\right\} \tag{34}$$

c is the number of classes, and the matrix $\mathbb{H}(\mathcal{A}_k) = [h_{ij}(\mathcal{A}_k)]_{c\times n}$ is termed as hypercuboid equivalence partition matrix of the condition attribute $\mathcal{A}_k$ [29], where

$$h_{ij}(\mathcal{A}_k) = \begin{cases} 1 & \text{if } \mathcal{L}_i \leqslant O_j(\mathcal{A}_k) \leqslant \mathcal{U}_i \\ 0 & \text{otherwise} \end{cases} \tag{35}$$

represents the membership of object O_j in the ith class β_i. Here $[\mathcal{L}_i, \mathcal{U}_i]$ represents the interval of ith class β_i according to the decision attribute set $\mathbb{D}$. The interval $[\mathcal{L}_i, \mathcal{U}_i]$ is the value range of condition attribute $\mathcal{A}_k$ with respect to class β_i. It is spanned by the objects with same class label β_i. That is, the value of each object O_j with class label β_i falls within interval $[\mathcal{L}_i, \mathcal{U}_i]$. A $c \times n$ hypercuboid equivalence partition matrix $\mathbb{H}(\mathcal{A}_k)$ represents the c-hypercuboid equivalence partitions of the universe generated by an equivalence relation. Each row of the matrix $\mathbb{H}(\mathcal{A}_k)$ is a hypercuboid equivalence class.

So, the relevance $\gamma_{\mathcal{A}_k}(\mathbb{D}) \in [0, 1]$. If $\gamma_{\mathcal{A}_k}(\mathbb{D}) = 1$, $\mathbb{D}$ depends totally on $\mathcal{A}_k$, if $0 < \gamma_{\mathcal{A}_k}(\mathbb{D}) < 1$, $\mathbb{D}$ depends partially on $\mathcal{A}_k$, and if $\gamma_{\mathcal{A}_k}(\mathbb{D}) = 0$, then $\mathbb{D}$ does not depend on $\mathcal{A}_k$. To calculate the significance of the feature $\mathcal{A}_k$ with respect to the set $\{\mathcal{A}_k, \mathcal{A}_l\}$ using (28), the joint relevance $\gamma_{\{\mathcal{A}_k,\mathcal{A}_l\}}(\mathbb{D})$ is to be computed. The joint relevance depends on the $c \times n$ hypercuboid equivalence partition matrix corresponding to the set $\{\mathcal{A}_k, \mathcal{A}_l\}$, which can be calculated from two $c \times n$ hypercuboid equivalence partition matrices $\mathbb{H}(\mathcal{A}_k)$ and $\mathbb{H}(\mathcal{A}_l)$ as follows:

$$\mathbb{H}(\{\mathcal{A}_k, \mathcal{A}_l\}) = \mathbb{H}(\mathcal{A}_k) \cap \mathbb{H}(\mathcal{A}_l) \tag{36}$$

$$\text{where} \quad h_{ij}(\{\mathcal{A}_k, \mathcal{A}_l\}) = h_{ij}(\mathcal{A}_k) \cap h_{ij}(\mathcal{A}_l). \tag{37}$$

D. Computational Complexity

Let $\mathcal{X}$ and $\mathcal{Y}$ be the two datasets with n samples and c classes, and p and q represent the number of features in $\mathcal{X}$ and $\mathcal{Y}$, respectively. Let us assume that the regularization parameters $\mathfrak{r}_x$ and $\mathfrak{r}_y$ have $\mathfrak{t}_x$ and $\mathfrak{t}_y$ possible values. Let $\mathcal{M} = \max(p, q)$ and $\mathcal{K} = \min(p, q)$, where the number of extracted features $\mathcal{D} << \mathcal{K}$. The computational complexity to calculate cross-covariance matrix C_{xy} is $\mathcal{O}(\mathcal{K}\mathcal{M}n)$, whereas that of covariance matrices C_{xx} and C_{yy} is $\mathcal{O}(\mathcal{K}^2 n + \mathcal{M}^2 n)$. In step 3, the eigenvalues Λ_x and Λ_y with corresponding eigenvectors Ψ_x and Ψ_y can be calculated with complexity $\mathcal{O}(\mathcal{K}^3 + \mathcal{M}^3)$ using Jacobi method. Hence, the total time complexity of these three steps is $(\mathcal{O}(\mathcal{K}\mathcal{M}n + \mathcal{K}^2 n + \mathcal{M}^2 n + \mathcal{K}^3 + \mathcal{M}^3) =)\mathcal{O}(\mathcal{M}^3)$. The step 4 has constant time complexity, which is $\mathcal{O}(1)$.

There is a loop in step 5, which is executed $(\mathfrak{t}_x \times \mathfrak{t}_y)$ times. The computational complexity to calculate $\Sigma\Sigma^T$ using (26) is $(\mathcal{O}(\mathcal{K}^3 + \mathcal{K}^2\mathcal{M} + \mathcal{K}\mathcal{M}^2 + \mathcal{M}^3) =)\mathcal{O}(\mathcal{M}^3)$. The computational complexity to calculate first $\mathcal{D}$ eigenvectors in step 5(b) is $\mathcal{O}(\mathcal{D}\mathcal{K}^2)$. Step 5(c) has complexity $\mathcal{O}(\mathcal{D}\mathcal{K}\mathcal{M})$. The computational complexity to calculate the basis vectors w_x and w_y is $\mathcal{O}(\mathcal{D}\mathcal{K}^2 + \mathcal{D}\mathcal{M}^2)$; and the canonical variables $\mathcal{U}$ and $\mathcal{V}$ have total $\mathcal{O}(\mathcal{D}\mathcal{K}n + \mathcal{D}\mathcal{M}n)$ time complexity. The computational complexity to extract first $\mathcal{D}$ features $\{\mathcal{A}\}$ is $\mathcal{O}(\mathcal{D}n)$. The time complexity to compute both relevance and significance of a feature is same, which is $\mathcal{O}(cn)$. In effect, the total complexity to compute both relevance and significance of $\mathcal{D}$ features is $\mathcal{O}(\mathcal{D}cn)$. Hence, step 5(f) has computational complexity $\mathcal{O}(\mathcal{D})$. Finally, step 5(g) has $\mathcal{O}(\mathcal{D}n)$ time complexity. Hence, the total complexity to execute the loop $(\mathfrak{t}_x \times \mathfrak{t}_y)$ times is $(\mathcal{O}(\mathfrak{t}_x\mathfrak{t}_y(\mathcal{M}^3 + \mathcal{D}\mathcal{K}^2 + \mathcal{D}\mathcal{K}\mathcal{M} + \mathcal{D}\mathcal{K}^2 + \mathcal{D}\mathcal{M}^2 + \mathcal{D}\mathcal{K}n + \mathcal{D}\mathcal{M}n + \mathcal{D}n + \mathcal{D}cn + \mathcal{D} + \mathcal{D}n)) =)\mathcal{O}(\mathfrak{t}_x\mathfrak{t}_y\mathcal{M}(\mathcal{M}^2 + \mathcal{D}n))$.

Hence, the overall computational complexity of the proposed algorithm to extract relevant and significant features, which are linearly correlated, is $(\mathcal{O}(\mathcal{M}^3 + \mathfrak{t}_x\mathfrak{t}_y\mathcal{M}(\mathcal{M}^2 + \mathcal{D}n)) =)\mathcal{O}(\mathfrak{t}_x\mathfrak{t}_y\mathcal{M}(\mathcal{M}^2 + \mathcal{D}n))$. On the other hand, the existing CCA, RCCA, and SRCCA algorithms have time complexity $\mathcal{O}(\mathcal{K}!)$, $\mathcal{O}(\mathfrak{t}_x\mathfrak{t}_y n\mathcal{K}!)$, and $\mathcal{O}(\mathfrak{t}_x\mathfrak{t}_y\mathcal{K}!)$, respectively, based on the analysis reported in [24].

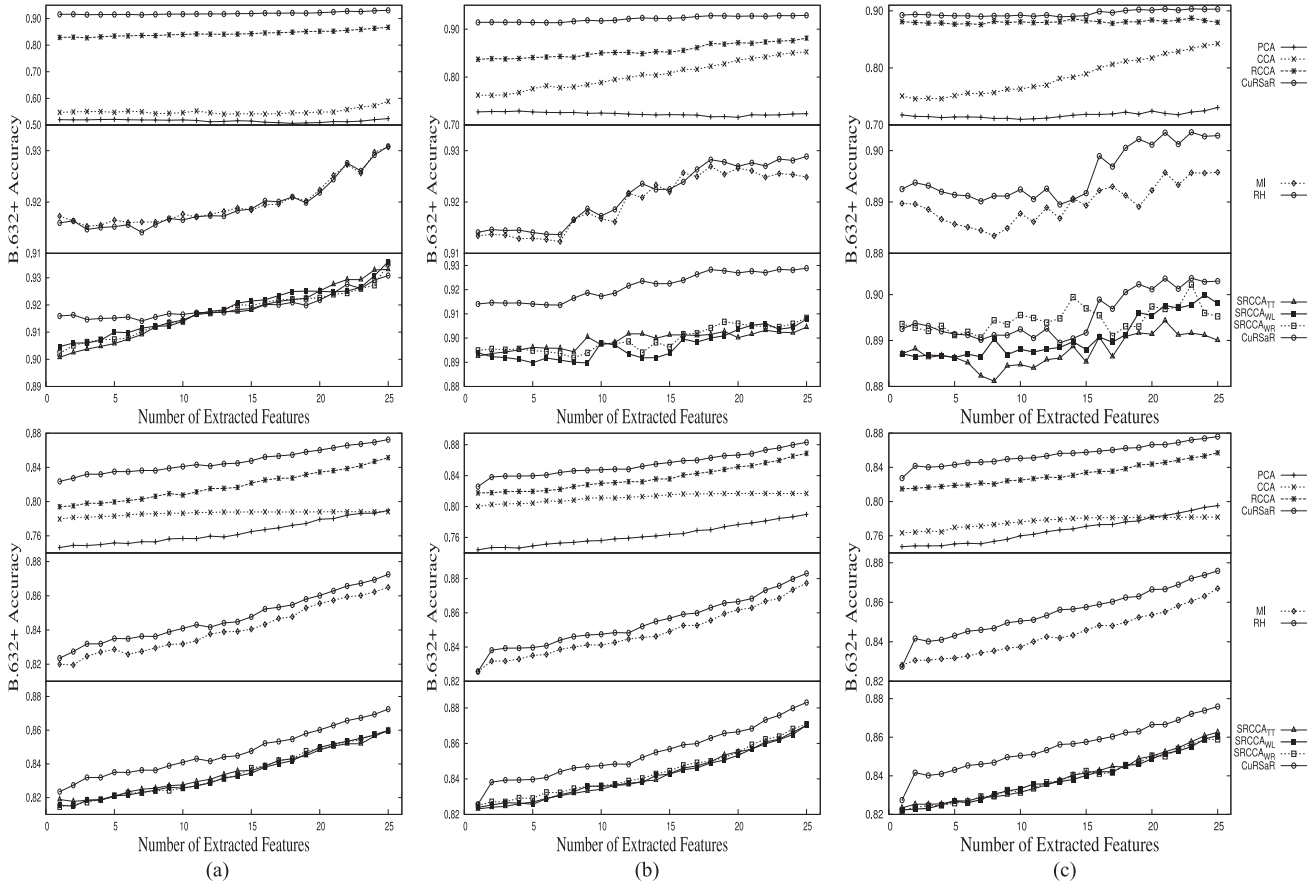

Fig. 1. Variation of $.632+$ bootstrap accuracy of the SVM over different number of extracted features (top: BRCA; bottom: OV). (a) Gene–DNA Methylation. (b) Gene–Protein. (c) Protein–DNA Methylation.

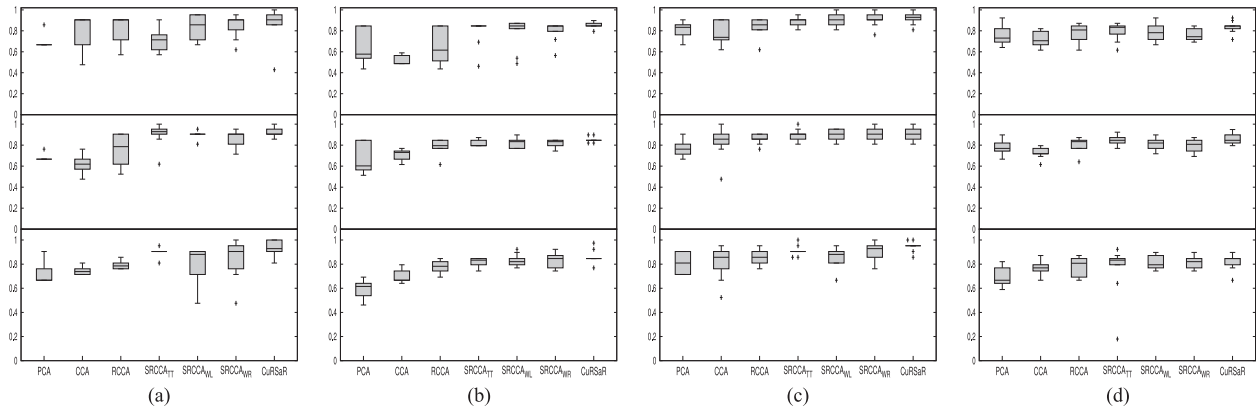

Fig. 2. Box and whisker plots for classification accuracy (top: gene–DNA methylation; middle: gene–protein; and bottom: protein–DNA methylation). (a) SVM on BRCA. (b) SVM on OV. (c) NNA on BRCA. (d) NNA on OV.

IV. Experimental Results and Discussion

The performance of the proposed feature extraction algorithm, termed as CuRSaR, is extensively studied and compared with that of some existing CCA based algorithms. The algorithms compared are principal component analysis (PCA), CCA, RCCA, several variants of SRCCA using t-test ($\mathrm{SRCCA_{TT}}$) [24], Wilcoxon rank sum test ($\mathrm{SRCCA_{WR}}$) [24], and Wilks's lambda test ($\mathrm{SRCCA_{WL}}$) [24]. The performance of rough hypercuboid approach is also compared with that of mutual information in the proposed feature extraction framework. All the algorithms are implemented in C language and run in Ubuntu 14.04 LTS having machine configuration Intel(R) Core(TM) i7-4790 CPU @ 3.60 GHz $\times 8$ and 32 GB RAM. The value of ω in (31) is set to 0.5, while $\mathfrak{r}_\chi$ and $\mathfrak{r}_y$ are varied within $[0.0, 1.0]$ with 0.1 as common difference. The source code of the CuRSaR algorithm and supplementary results are available at www.isical.ac.in/$\sim$ bibl/results/cursar/cursar.html.

A. Description of Datasets

Two multimodal datasets, namely, BRCA and OV, are used in the current research work, each having three different modalities, namely, gene expression, protein expression, and DNA methylation. These datasets are downloaded from TCGA. The BRCA dataset contains a total number of 204 breast invasive carcinoma samples, classified into two classes: 189 samples of infiltrating ductal carcinoma and 15 samples of infiltrating lobular carcinoma. On the other hand, OV dataset consists of 379 ovarian serous cystadenocarcinoma samples, grouped into two categories: 51 samples of grade 2 and 328 samples of grade 3 ovarian serous cystadenocarcinoma. Both datasets contain expressions of 17 814 genes and β values of 27 578 methylated DNAs. While BRCA data has expression of 142 proteins, OV data has expression of 222 proteins. In the current study, 2000 top-ranked features, based on their variances, are taken from both gene and methylation data.

B. Classifiers Used

To evaluate the performance of different algorithms, both support vector machine (SVM) [36] and nearest neighbor algorithm (NNA) [37] are used in the current study. Being a margin classifier, the SVM defines the boundary between data samples of different classes by drawing an optimal hyperplane. The hyperplane leads to good generalization properties as it maximizes the margin between different classes. In the current work, linear kernels are used. On the other hand, the NNA [37] is used for evaluating the effectiveness of the generated feature set for classification. It classifies samples based on closest training samples in the feature space. The class label of a sample is assigned by the class label of its closest neighbor.

C. Experimental Setup

Both .632+ bootstrap method [38] and ten-fold cross-validation [37] are performed to compute the classification accuracy and F1 score of different approaches. In order to minimize the variability and biasedness of derived results, the $B.632+$ bootstrap approach [38] is used, while ten-fold cross-validation is performed to analyze the statistical significance of the derived results. In each of these approaches, a set of correlated features is first generated for each training set, and then both SVM and NNA are trained with this feature set. After the training, the information of correlated features that were selected for the training set is used to generate test set and then the class label of the test sample is predicted using the SVM and NNA. For each dataset, 25 top-ranked correlated features are selected for the analysis.

Fig. 1 presents the variation of .632+ bootstrap ($B.632+$) accuracy of the SVM considering 100 bootstrap samples. Results are reported for 25 top-ranked extracted features. The comparative performance analysis of different algorithms are also studied using box and whisker plots, tables of means, standard deviations, and p-values computed through both paired-t (one-tailed) and Wilcoxon signed-rank (one-tailed) tests. The ten-fold cross-validation is performed on each pair of modalities for this analysis. Figs. 2 and 3 show the box and whisker plots for classification accuracy and F1 score, respectively, on six pair of modalities. In these plots, the central line on each box is the median, the upper and lower boundaries of the box are 25th and 75th percentiles, that is, upper quartile and lower quartile, respectively, and the whiskers extend to three standard deviations from mean to include most extreme data points. The outliers are marked as '+', plotted individually. The outlier represents the test set for which a method produces classification accuracy or F1 score much worse than for the other test sets of the same dataset.

Table I presents the classification accuracy and F1 score for PCA on three individual modalities as a baseline comparison. On the other hand, Tables II–V report the means and standard deviations of accuracy and F1 score for all the methods. The computed p-values are also reported in these tables with respect to the proposed method for two datasets and three pairs of modalities. The best mean values are marked in bold in these tables.

D. Comparative Performance Analysis

This section presents the performance of the proposed CuRSaR algorithm on three pairs of modalities, namely, gene–protein, gene–DNA methylation, and protein–DNA methylation, of two datasets, namely, BRCA and OV. From the results reported in Fig. 1, it is seen that the $B.632+$ accuracy of the SVM for the proposed method increases as the number of extracted features increases. Also, the $B.632+$ accuracy of the CuRSaR algorithm is higher as compared to existing methods. All the results, presented in Figs. 2 and 3, and Tables II–V, establish the fact that the proposed method attains best mean classification accuracy in all the cases, irrespective of the datasets, pairs of modalities, and classifiers used. Also, the proposed CuRSaR achieves best mean F1 score in 11 cases out of total 12 cases.

The results reported in Fig. 1 show that the $B.632+$ accuracy of the CuRSaR is significantly higher as compared to existing PCA, CCA,and RCCA, irrespective of the pair of modalities, datasets, and number of extracted features. Moreover, it can be seen from the results, reported in Figs. 2 and 3, and Tables II–V using ten-fold cross-validation, that the proposed CuRSaR performs significantly better than PCA, CCA, and RCCA in 48, 37, and 32 cases, respectively, out of total 48 cases each, considering 0.05 as the level of significance. For remaining cases, the performance of the CuRSaR is better than both CCA and RCCA, but not significantly. Also, the performance of the PCA is significantly poor compared to CCA, RCCA, and CuRSaR, even considering individual modality as reported in Table I. This is mainly due to the drastic variation and noisy nature of different modalities. The significantly better performance of the proposed algorithm over these three approaches is achieved due to the fact that the CuRSaR algorithm extracts features by maximizing the relevance and significance of the features. Both relevance and significance measures depend on the information of sample categories. On the other hand, PCA, CCA, and RCCA extract features from two different modalities without considering the supervised information of class labels. In effect, the proposed algorithm is able to extract more relevant and significant features from a pair of modalities.

In the proposed CuRSaR method, both relevance and significance of an extracted feature are calculated based on the theory of hypercuboid equivalence partition matrix of rough hypercuboid approach. The relevance of a feature with respect to the class labels is calculated using (32), while the significance of a feature with respect to the already-extracted features is

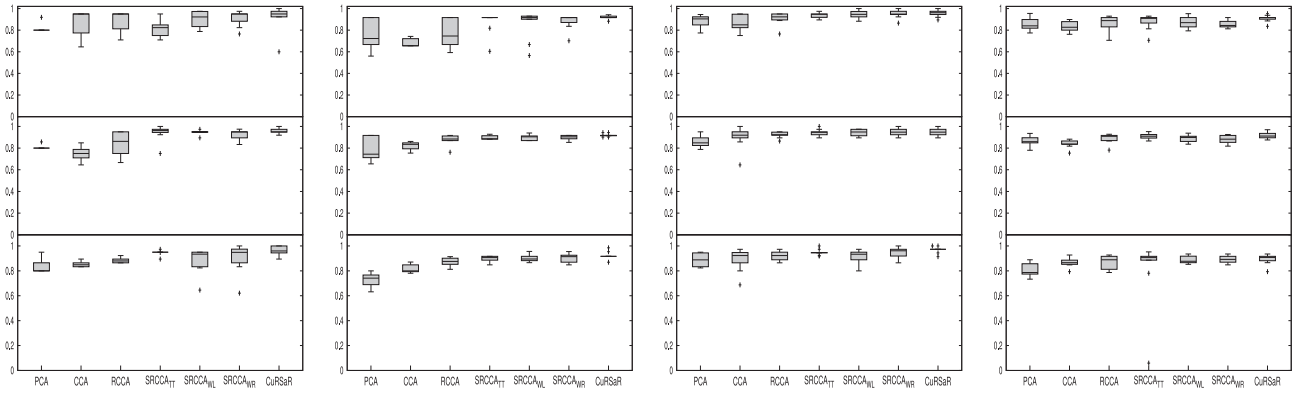

Fig. 3. Box and whisker plots for F1 score (top: gene–DNA methylation; middle: gene–protein; and bottom: protein–DNA methylation). (a) SVM on BRCA. (b) SVM on OV. (c) NNA on BRCA. (d) NNA on OV.

TABLE I
CLASSIFICATION ACCURACY AND F1 SCORE OF SVM AND NNA FOR PCA

Different datasets	Different classifiers	Different statistics	Gene		DNA methylation		Protein	
			Accuracy	F1 Score	Accuracy	F1 Score	Accuracy	F1 Score
BRCA	SVM	Mean	0.685714	0.812973	0.676190	0.806486	0.685714	0.811892
		StdDev	0.040156	0.027349	0.030117	0.020512	0.060234	0.037605
	NNA	Mean	0.823810	0.896375	0.776190	0.870441	0.814286	0.884599
		StdDev	0.089932	0.058214	0.089932	0.057333	0.137373	0.088070
OV	SVM	Mean	0.700000	0.808460	0.666667	0.777771	0.682051	0.795386
		StdDev	0.157158	0.116881	0.161265	0.128819	0.178139	0.132979
	NNA	Mean	0.753846	0.848403	0.766667	0.861755	0.782051	0.868051
		StdDev	0.101274	0.066739	0.072976	0.048232	0.085683	0.055290

TABLE II
CLASSIFICATION ACCURACY OF THE SVM FOR CuRSaR AND OTHER METHODS

Datasets	Different algorithms	Gene–protein				Protein–DNA methylation				Gene–DNA methylation			
		Mean	StdDev	Wilcoxon:p	Paired-t:p	Mean	StdDev	Wilcoxon:p	Paired-t:p	Mean	StdDev	Wilcoxon:p	Paired-t:p
BRCA	PCA	0.676190	0.030117	2.24E–03	6.90E–08	0.709524	0.079206	3.68E–03	6.22E–05	0.685714	0.060234	1.80E–02	3.68E–03
	CCA	0.623810	0.082326	2.47E–03	5.58E–06	0.747619	0.039203	2.50E–03	2.01E–05	0.785714	0.162262	1.31E–01	1.79E–01
	RCCA	0.752381	0.164651	8.58E–03	5.79E–03	0.790476	0.033296	3.71E–03	2.46E–04	0.814286	0.131756	1.63E–01	2.50E–01
	SRCCA$_{TT}$	0.900000	0.106361	4.44E–01	2.73E–01	0.900000	0.035136	6.08E–02	9.67E–02	0.704762	0.104810	2.95E–02	1.37E–02
	SRCCA$_{WL}$	0.900000	0.035136	5.12E–02	8.39E–02	0.809524	0.140187	2.19E–02	3.38E–02	0.833333	0.112799	2.20E–01	3.18E–01
	SRCCA$_{WR}$	0.876190	0.075125	5.12E–02	6.71E–02	0.852381	0.162573	1.42E–01	1.10E–01	0.852381	0.108704	2.64E–01	4.14E–01
	SRCCA$_{RSMI}$	0.914286	0.049181	5.35E–01	3.63E–01	0.895238	0.077111	1.18E–01	1.05E–01	0.852381	0.069007	1.29E–01	4.06E–01
	CuRSaR	**0.919048**	0.039203			**0.928571**	0.064476			**0.866667**	0.159995		
OV	PCA	0.676923	0.149120	8.58E–03	2.59E–03	0.594872	0.079996	2.50E–03	4.61E–06	0.653846	0.170619	8.98E–03	3.16E–03
	CCA	0.715385	0.050492	2.49E–03	3.80E–05	0.694872	0.054661	2.47E–03	6.91E–05	0.515385	0.045948	2.45E–03	7.22E–09
	RCCA	0.787179	0.068429	1.02E–02	9.95E–03	0.774359	0.052409	5.66E–03	1.06E–03	0.666667	0.163960	5.71E–03	2.65E–03
	SRCCA$_{TT}$	0.820513	0.034188	1.36E–02	1.19E–02	0.812821	0.041959	3.03E–02	2.55E–02	0.792308	0.125876	5.74E–02	8.68E–02
	SRCCA$_{WL}$	0.823077	0.045948	2.53E–02	1.59E–02	0.828205	0.051353	2.00E–01	1.39E–01	0.787179	0.146076	1.68E–01	9.52E–02
	SRCCA$_{WR}$	0.812821	0.041959	2.32E–02	3.09E–02	0.828205	0.061693	1.14E–01	1.35E–01	0.797436	0.091617	2.11E–02	5.66E–02
	SRCCA$_{RSMI}$	0.843590	0.030698	1.67E–01	9.67E–02	0.846154	0.029608	2.75E–01	2.65E–01	0.817949	0.100004	1.79E–01	1.54E–01
	CuRSaR	**0.851282**	0.026482			**0.858974**	0.054393			**0.851282**	0.026482		

computed using (28). However, other measures, such as mutual information can also be used to compute both relevance and significance of a feature [28], [39], [40]. In order to establish the importance of rough hypercuboid (RH) approach over mutual information (MI), extensive experimental results are reported in Fig. 1 for three pairs of modalities of two datasets. Subsequent discussions analyze the results with respect to the $B.632+$ accuracy of the SVM. All the results reported in Fig. 1 confirm that the performance of hypercuboid equivalence partition matrix is better than that of mutual information in most of the cases, irrespective of the pairs of modalities and datasets used. Also, the mean values of both classification accuracy and F1 score obtained using mutual information (SRCCA$_{RSMI}$) for ten-fold cross-validation, as reported in Tables II–V, are lower than

TABLE III
CLASSIFICATION ACCURACY OF THE NNA FOR CuRSaR AND OTHER METHODS

Datasets	Different algorithms	Gene–protein				Protein–DNA methylation				Gene–DNA methylation			
		Mean	StdDev	Wilcoxon:p	Paired-t:p	Mean	StdDev	Wilcoxon:p	Paired-t:p	Mean	StdDev	Wilcoxon:p	Paired-t:p
BRCA	PCA	0.771429	0.073771	3.26E–03	2.46E–04	0.809524	0.080937	3.68E–03	2.46E–04	0.809524	0.074451	5.14E–03	3.38E–03
	CCA	0.833333	0.144174	1.03E–01	1.03E–01	0.814286	0.131756	5.66E–03	5.97E–03	0.771429	0.104810	3.82E–03	6.48E–04
	RCCA	0.857143	0.044896	4.04E–02	4.66E–02	0.852381	0.061271	6.31E–03	3.56E–03	0.838095	0.087518	8.78E–03	8.16E–03
	$\mathrm{SRCCA}_{\mathrm{TT}}$	0.900000	0.052405	2.31E–01	2.78E–01	0.909524	0.041695	1.26E–02	2.64E–02	0.890476	0.045175	4.42E–02	4.43E–02
	$\mathrm{SRCCA}_{\mathrm{WL}}$	0.890476	0.055214	3.06E–01	2.67E–01	0.857143	0.080937	5.86E–03	4.24E–03	0.904762	0.063492	1.43E–01	1.55E–01
	$\mathrm{SRCCA}_{\mathrm{WR}}$	0.904762	0.054986	4.57E–01	3.79E–01	0.904762	0.074451	3.54E–02	3.38E–02	0.914286	0.070273	2.90E–01	2.78E–01
	$\mathrm{SRCCA}_{\mathrm{RSMI}}$	0.890476	0.081092	2.46E–01	1.87E–01	0.919048	0.055214	1.03E–01	9.67E–02	0.914286	0.070273	3.27E–01	2.78E–01
	CuRSaR	**0.909524**	0.061271			**0.947619**	0.041695			**0.923810**	0.060234		
OV	PCA	0.676923	0.149120	3.74E–03	1.72E–04	0.692308	0.075485	2.49E–03	1.11E–04	0.758974	0.089805	2.49E–02	1.71E–02
	CCA	0.725641	0.049910	2.50E–03	6.39E–04	0.771795	0.053308	5.08E–02	4.87E–02	0.725641	0.071560	3.68E–03	1.76E–04
	RCCA	0.807692	0.067570	7.03E–02	6.39E–02	0.782051	0.083086	6.38E–02	4.79E–02	0.782051	0.083086	5.38E–02	6.00E–02
	$\mathrm{SRCCA}_{\mathrm{TT}}$	0.841026	0.049543	3.36E–01	2.57E–01	0.758974	0.216292	4.76E–01	2.23E–01	0.797436	0.082380	1.03E–01	1.18E–01
	$\mathrm{SRCCA}_{\mathrm{WL}}$	0.810256	0.065316	3.97E–02	5.63E–02	0.812821	0.051353	2.53E–01	2.80E–01	0.789744	0.091097	3.91E–02	3.20E–02
	$\mathrm{SRCCA}_{\mathrm{WR}}$	0.797436	0.059768	2.03E–02	1.79E–02	0.817949	0.051919	3.63E–01	4.60E–01	0.761538	0.056759	5.81E–03	6.30E–04
	$\mathrm{SRCCA}_{\mathrm{RSMI}}$	0.828205	0.049910	1.70E–01	1.14E–01	0.802564	0.038319	1.46E–01	2.14E–01	0.797436	0.074951	1.36E–02	2.23E–02
	CuRSaR	**0.853846**	0.049910			**0.820513**	0.065092			**0.838462**	0.055457		

TABLE IV
F1 SCORE OF THE SVM FOR CuRSaR AND OTHER METHODS

Datasets	Different algorithms	Gene–protein				Protein–DNA methylation				Gene–DNA methylation			
		Mean	StdDev	Wilcoxon:p	Paired-t:p	Mean	StdDev	Wilcoxon:p	Paired-t:p	Mean	StdDev	Wilcoxon:p	Paired-t:p
BRCA	PCA	0.805714	0.018070	2.46E–03	3.13E–08	0.827973	0.050651	3.40E–03	6.01E–05	0.811892	0.037605	3.67E–02	8.78E–03
	CCA	0.752934	0.063423	2.52E–03	6.39E–06	0.855073	0.025380	2.52E–03	2.06E–05	0.867194	0.113472	1.63E–01	1.98E–01
	RCCA	0.840054	0.119284	1.41E–02	7.09E–03	0.882635	0.020549	3.74E–03	2.84E–04	0.885453	0.094030	1.63E–01	2.69E–01
	$\mathrm{SRCCA}_{\mathrm{TT}}$	0.940524	0.070368	2.00E–01	2.30E–01	0.946508	0.019626	1.66E–01	1.35E–01	0.812500	0.076282	2.96E–02	2.01E–02
	$\mathrm{SRCCA}_{\mathrm{WL}}$	0.946245	0.019592	1.99E–01	1.08E–01	0.886568	0.097932	4.63E–02	4.20E–02	0.901693	0.071398	1.57E–01	3.58E–01
	$\mathrm{SRCCA}_{\mathrm{WR}}$	0.930903	0.044582	1.18E–01	7.14E–02	0.908667	0.115023	1.42E–01	1.16E–01	0.914100	0.069811	2.23E–01	4.54E–01
	$\mathrm{SRCCA}_{\mathrm{RSMI}}$	0.953025	0.026219	2.48E–01	3.57E–01	0.942603	0.043491	1.64E–01	1.21E–01	0.913560	0.042475	5.70E–02	4.45E–01
	CuRSaR	**0.955836**	0.021775			**0.960002**	0.035594			**0.919487**	0.114662		
OV	PCA	0.793207	0.109827	1.04E–02	2.91E–03	0.730561	0.057004	2.53E–03	2.22E–06	0.765392	0.139100	8.98E–03	4.00E–03
	CCA	0.823437	0.035102	2.53E–03	2.33E–05	0.813724	0.033526	2.52E–03	2.36E–05	0.679139	0.038996	2.45E–03	1.42E–08
	RCCA	0.877994	0.044900	7.53E–03	9.94E–03	0.870792	0.034335	3.84E–03	1.09E–03	0.775962	0.130273	5.81E–03	3.15E–03
	$\mathrm{SRCCA}_{\mathrm{TT}}$	0.897980	0.019315	6.23E–03	6.44E–03	0.895236	0.026597	2.48E–02	2.52E–02	0.875529	0.100376	8.59E–02	1.06E–01
	$\mathrm{SRCCA}_{\mathrm{WL}}$	0.899035	0.024790	8.30E–03	4.93E–03	0.901851	0.029458	1.07E–01	1.04E–01	0.860160	0.131228	1.73E–01	9.54E–02
	$\mathrm{SRCCA}_{\mathrm{WR}}$	0.894672	0.024966	1.79E–02	2.54E–02	0.901083	0.036584	6.92E–02	9.24E–02	0.880995	0.068573	1.36E–02	6.29E–02
	$\mathrm{SRCCA}_{\mathrm{RSMI}}$	0.913909	0.017305	5.79E–02	9.30E–02	0.915592	0.017361	3.05E–01	2.67E–01	0.892488	0.080014	2.33E–01	1.64E–01
	CuRSaR	**0.918377**	0.014219			**0.922515**	0.029699			**0.918437**	0.015511		

that obtained using rough hypercuboid approach (CuRSaR), irrespective of the datasets, pairs of modalities, and classifiers used. The proposed method with rough hypercuboid approach (CuRSaR) achieves significantly better p-values than that with mutual information ($\mathrm{SRCCA}_{\mathrm{RSMI}}$) in four cases, better but not significant in 19 cases, and worse but not significant in one case, out of total 24 cases. The better performance of the rough hypercuboid approach is achieved due to the fact that the hypercuboid equivalence partition matrix evaluates the quality of an extracted feature set through supervised granulation process that utilizes class information of samples. Also, it provides an efficient way to calculate degree of dependency of class labels on feature set in approximation spaces. In effect, a reduced set of features having maximum relevance and significance is being obtained using the proposed method.

Finally, the performance of the proposed CuRSaR algorithm is compared with that of several existing SRCCA algorithms, namely, $\mathrm{SRCCA}_{\mathrm{TT}}$ [24], $\mathrm{SRCCA}_{\mathrm{WR}}$ [24], and $\mathrm{SRCCA}_{\mathrm{WL}}$ [24]. Comparative results are reported in Fig. 1 for different number of extracted features considering three pairs of modalities and two datasets. From the results reported in Fig. 1, it is seen that the performance of the CuRSaR algorithm is better than that of other SRCCA algorithms in almost all the cases with respect to the $B.632+$ accuracy of the SVM, irrespective of the number of extracted features, datasets, and pairs of modalities used. Also, the box and whiskers plots presented in Figs. 2 and 3 and mean values of accuracy and F1 score reported in Tables II–V show that the proposed method attains the best mean values in 11 cases, while $\mathrm{SRCCA}_{\mathrm{WR}}$ achieves it in only one case for F1 score using the NNA. Moreover, the proposed CuRSaR attains significantly better p-values than the existing SRCCA algorithms in 53 cases, better but not significant in 89 cases, and worse but not significant in two cases, out of total 144 cases. Finally, Table VI compares the execution time of different algorithms, which shows that the execution time of the proposed RCCA formulation, introduced in Section III-A, is significantly

TABLE V
F1 SCORE OF THE NNA FOR CuRSaR AND OTHER METHODS

Datasets	Different algorithms	Gene–protein				Protein–DNA methylation				Gene–DNA methylation			
		Mean	StdDev	Wilcoxon:p	Paired-t:p	Mean	StdDev	Wilcoxon:p	Paired-t:p	Mean	StdDev	Wilcoxon:p	Paired-t:p
BRCA	PCA	0.860676	0.050960	3.46E–03	1.58E–04	0.889551	0.050057	3.82E–03	2.94E–04	0.886223	0.052737	6.26E–03	4.07E–03
	CCA	0.900350	0.098939	1.31E–01	1.08E–01	0.890804	0.087358	5.40E–03	1.04E–02	0.863473	0.069560	3.46E–03	9.19E–04
	RCCA	0.921820	0.026317	5.67E–02	5.21E–02	0.918390	0.036003	1.04E–02	5.90E–03	0.908659	0.056315	1.04E–02	1.22E–02
	$\mathrm{SRCCA}_{\mathrm{TT}}$	0.944596	0.028962	2.33E–01	2.83E–01	0.949333	0.023103	6.27E–02	3.51E–02	0.939279	0.025870	4.63E–02	5.49E–02
	$\mathrm{SRCCA}_{\mathrm{WL}}$	0.939590	0.031173	3.90E–01	2.78E–01	0.917510	0.050039	7.19E–03	5.91E–03	0.946637	0.036823	3.37E–01	1.69E–01
	$\mathrm{SRCCA}_{\mathrm{WR}}$	0.947295	0.030468	3.37E–01	3.89E–01	0.947041	0.041888	5.74E–02	4.03E–02	0.952153	0.039604	2.50E–01	2.78E–01
	$\mathrm{SRCCA}_{\mathrm{RSMI}}$	0.937895	0.046521	1.04E–01	1.69E–01	0.954582	0.030819	1.24E–01	1.14E–01	0.950951	0.041284	2.97E–01	2.56E–01
	CuRSaR	**0.949889**	0.033858			**0.969657**	0.024964			**0.957577**	0.033554		
OV	PCA	0.793207	0.109827	3.84E–03	4.78E–04	0.805719	0.051850	2.53E–03	1.41E–04	0.853425	0.056873	1.42E–02	1.09E–02
	CCA	0.838164	0.035418	2.53E–03	1.20E–03	0.867218	0.035212	8.44E–02	9.38E–02	0.836694	0.046084	2.53E–03	2.26E–04
	RCCA	0.891351	0.043883	1.57E–01	1.17E–01	0.870083	0.056013	6.41E–02	7.48E–02	0.867081	0.068967	4.55E–02	6.88E–02
	$\mathrm{SRCCA}_{\mathrm{TT}}$	0.907589	0.027493	4.30E–01	3.08E–01	0.813962	0.269150	5.20E–01	2.04E–01	0.878312	0.069763	1.24E–01	1.31E–01
	$\mathrm{SRCCA}_{\mathrm{WL}}$	0.887372	0.037076	4.29E–02	5.27E–02	0.888822	0.028168	2.38E–01	3.54E–01	0.873979	0.056436	3.98E–02	2.76E–02
	$\mathrm{SRCCA}_{\mathrm{WR}}$	0.880799	0.035917	2.13E–02	1.72E–02	**0.892835**	0.030058	3.90E–01	5.13E–01	0.856286	0.037255	5.86E–03	5.12E–04
	$\mathrm{SRCCA}_{\mathrm{RSMI}}$	0.898963	0.028031	1.73E–01	1.17E–01	0.882230	0.023692	1.18E–01	2.58E–01	0.877546	0.045696	1.39E–02	1.51E–02
	CuRSaR	**0.912948**	0.028749			0.892284	0.040470			**0.908272**	0.031468		

TABLE VI
EXECUTION TIME (IN MILLI SECOND) OF CuRSaR AND OTHER METHODS

Different algorithms	Gene–protein		Protein–DNA methylation		Gene–DNA methylation	
	BRCA	OV	BRCA	OV	BRCA	OV
PCA	34 339	45 784	32 026	45 875	163 162	211 391
CCA	96 819	117 163	84 738	100 129	527 037	631 754
RCCA: Existing	11 715 222	14 176 858	10 253 552	12 115 821	63 771 641	76 442 414
RCCA: Proposed	6 832 146	8 371 641	5 019 576	7 399 362	44 735 195	46 289 456
$\mathrm{SRCCA}_{\mathrm{TT}}$	11 716 075	14 178 912	10 254 051	12 117 724	63 772 309	76 444 295
$\mathrm{SRCCA}_{\mathrm{WL}}$	11 716 099	14 178 931	10 254 369	12 117 736	63 772 606	76 444 359
$\mathrm{SRCCA}_{\mathrm{WR}}$	11 716 031	14 178 864	10 254 280	12 117 707	63 772 492	76 444 492
CuRSaR	10 470 714	12 923 372	6 522 378	10 238 856	55 230 255	74 740 355

lower than that of existing RCCA. Also, the time required for the proposed CuRSaR algorithm is lower as compared to existing RCCA and different variants of SRCCA; although both PCA and CCA take lesser time compared to the proposed CuRSaR as expected.

The existing SRCCA algorithms consider only the correlation of first pair of canonical variables [24]. In effect, other canonical variable pairs may have insignificant relation with the first pair of canonical variables or may introduce some irrelevant features in the whole extracted feature set, which may degrade the prediction capability of the classifiers used. Also, the existing SRCCA algorithms fail to address the problem of uncertainty associated with omics data analysis. On the other hand, the proposed algorithm considers both relevance and significance measures of all extracted features while optimizing the regularization parameters. The rough hypercuboid approach, employed in proposed algorithm, can also efficiently handle the uncertainty due to imprecision in computation and vagueness in class definition. In effect, the proposed method provides significantly better results as compared to existing algorithms in most of the cases. Moreover, the analytical formulation introduced in this paper makes the computational complexity of the proposed CuRSaR algorithm significantly lower than existing RCCA and SRCCA.

V. CONCLUSION

This paper presents a new feature extraction algorithm, termed as CuRSaR, from two multidimensional datasets. The merits of CCA and rough sets are integrated judiciously to develop the proposed method. To establish the relation between regularization parameters and CCA, a theoretical formulation is presented. It helps the proposed CuRSaR algorithm to extract required number of correlated features, which are relevant with respect to class label and significant among them. The hypercuboid equivalence partition matrix is used to compute both relevance and significance of a feature. The optimum regularization parameters of CCA are determined using the equivalence partition matrix. The effectiveness of the proposed algorithm, along with a comparison with other algorithms, has been demonstrated considering three different modalities, namely, gene expression, protein expression, and DNA methylation. The concept of hypercuboid equivalence partition matrix is found to be successful in extracting relevant and significant features from multimodal high dimensional real life datasets.

REFERENCES

[1] H. Hotelling, "Relations between two sets of variates," *Biometrika*, vol. 28, no. 3/4, pp. 321–377, 1936.

[2] Y. Yamanishi *et al.*, "Protein network inference from multiple genomic data: A supervised approach," *Bioinformatics*, vol. 20, pp. i363–i370, 2004.
[3] Z. Lin *et al.*, "Frequency recognition based on canonical correlation analysis for SSVEP-Based BCIs," *IEEE Trans. Biomed. Eng.*, vol. 53, no. 12, pp. 2610–2614, Dec. 2006.
[4] M. Li *et al.*, "OI and fMRI signal separation using both temporal and spatial autocorrelations," *IEEE Trans. Biomed. Eng.*, vol. 57, no. 8, pp. 1917–1926, Aug. 2010.
[5] G. R. Wu *et al.*, "Multiscale causal connectivity analysis by canonical correlation: Theory and application to epileptic brain," *IEEE Trans. Biomed. Eng.*, vol. 58, no. 11, pp. 3088–3096, Nov. 2011.
[6] M. Hassan *et al.*, "Combination of canonical correlation analysis and empirical mode decomposition applied to denoising the labor electrohysterogram," *IEEE Trans. Biomed. Eng.*, vol. 58, no. 9, pp. 2441–2447, Sep. 2011.
[7] K. T. Sweeney *et al.*, "The use of ensemble empirical mode decomposition with canonical correlation analysis as a novel artifact removal technique," *IEEE Trans. Biomed. Eng.*, vol. 60, no. 1, pp. 97–105, Jan. 2013.
[8] C. Soneson *et al.*, "Integrative analysis of gene expression and copy number alterations using canonical correlation analysis," *BMC Bioinformat.*, vol. 11, no. 1, p. 191, 2010.
[9] K. A. L. Cao *et al.*, "integrOmics: An R package to unravel relationships between two omics datasets," *Bioinformatics*, vol. 25, no. 21, pp. 2855–2856, 2009.
[10] Y. Yamanishi *et al.*, "Supervised enzyme network inference from the integration of genomic data and chemical information," *Bioinformatics*, vol. 21, pp. i468–i477, 2005.
[11] D. Lin *et al.*, "Group sparse canonical correlation analysis for genomic data integration," *BMC Bioinformat.*, vol. 14, no. 1, p. 245, 2013.
[12] K. A. L. Cao *et al.*, "Sparse canonical methods for biological data integration: Application to a cross-platform study," *BMC Bioinformat.*, vol. 10, no. 1, pp. 1–17, 2009.
[13] S. A. Rifkin and J. Kim, "Geometry of gene expression dynamics," *Bioinformatics*, vol. 18, no. 9, pp. 1176–1183, 2002.
[14] L. J. Revell and A. S. Harrison, "PCCA: A program for phylogenetic canonical correlation analysis," *Bioinformatics*, vol. 24, no. 7, pp. 1018–1020, 2008.
[15] Z. Gou and C. Fyfe, "A canonical correlation neural network for multicollinearity and functional data," *Neural Netw.*, vol. 17, no. 2, pp. 285–293, 2004.
[16] I. Gonzalez *et al.*, "Highlighting relationships between heterogeneous biological data through graphical displays based on regularized canonical correlation analysis," *J. Biological Syst.*, vol. 17, no. 2, pp. 173–199, 2009.
[17] H. D. Vinod, "Canonical ridge and econometrics of joint production," *J. Econometrics*, vol. 4, no. 2, pp. 147–166, 1976.
[18] I. Gonzalez *et al.*, "CCA: An R package to extend canonical correlation analysis," *J. Statist. Softw.*, vol. 23, no. 12, pp. 1–14, 2008.
[19] T. D. Bie and B. D. Moor, "On the regularization of canonical correlation analysis," in *Proc. 4th Int. Symp. Ind. Compon. Anal. Blind Signal Separation*, 2003, pp. 785–790.
[20] A. E. Hoerl and R. W. Kennard, "Ridge regression: Biased estimation for nonorthogonal problems," *Technometrics*, vol. 12, no. 1, pp. 55–67, 1970.
[21] R. Cruz-Cano and M.-L. T. Lee, "Fast regularized canonical correlation analysis," *Comput. Statist. Data Anal.*, vol. 70, pp. 88–100, 2014.
[22] L. An *et al.*, "Person Re-Identification by robust canonical correlation analysis," *IEEE Signal Process. Lett.*, vol. 22, no. 8, pp. 1103–1107, Aug. 2015.
[23] M. Kang *et al.*, "eQTL mapping study via regularized sparse canonical correlation analysis," in *Proc. 12th Int. Conf. Mach. Learn. Appl.*, 2013, pp. 129–134.
[24] A. Golugula *et al.*, "Supervised regularized canonical correlation analysis: Integrating histologic and proteomic measurements for predicting biochemical recurrence following prostate surgery," *BMC Bioinformat.*, vol. 12, no. 1, p. 483, 2011.
[25] G. Lee *et al.*, "Supervised multi-view canonical correlation analysis (sMVCCA): Integrating histologic and proteomic features for predicting recurrent prostate cancer," *IEEE Trans. Med. Imag.*, vol. 34, no. 1, pp. 284–297, Jan. 2015.
[26] P. Maji and A. Mandal, "Rough hypercuboid based supervised regularized canonical correlation for multimodal data analysis," *Fundamenta Informaticae*, vol. 148, no. 1–2, pp. 133–155, 2016.
[27] P. Maji and S. K. Pal, *Rough-Fuzzy Pattern Recognition: Applications in Bioinformatics and Medical Imaging*. Hoboken, NJ, USA: Wiley, 2012.
[28] P. Maji and S. Paul, "Rough set based maximum relevance-maximum significance criterion and gene selection from microarray data," *Int. J. Approximate Reasoning*, vol. 52, no. 3, pp. 408–426, 2011.
[29] P. Maji, "Rough hypercuboid approach for feature selection in approximation spaces," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 1, pp. 16–29, Jan. 2014.
[30] S. Paul and P. Maji, "μHEM for identification of differentially expressed miRNAs using hypercuboid equivalence partition matrix," *BMC Bioinformat.*, vol. 14, no. 1, 2013, Art. no. 266.
[31] Y. Guo *et al.*, "Regularized linear discriminant analysis and its application in microarrays," *Biostatistics*, vol. 8, no. 1, pp. 86–100, 2007.
[32] N. H. Timm *et al.*, *Applied Multivariate Analysis*. New York, NY, USA: Springer, 2002.
[33] E. J. Ientilucci, "Using the singular value decomposition," Chester F. Carlson Center Imag. Sci., Rochester Inst. Technol., Rochester, NY, YSA, Tech. Rep., Version no. v1.0, 2003.
[34] G. M. L. Gladwell, "On isospectral spring - mass systems," *Inverse Problems*, vol. 11, no. 3, pp. 591–602, 1995.
[35] G. H. Golub and C. F. V. Loan, *Matrix Computations*, 3rd ed. Baltimore, MD, USA: The Johns Hopkins Univ. Press, 1996.
[36] V. Vapnik, *The Nature of Statistical Learning Theory*. New York, NY, USA: Springer, 1995.
[37] R. O. Duda *et al.*, *Pattern Classification and Scene Analysis*. New York, NY, USA: Wiley, 1999.
[38] B. Efron and R. Tibshirani, "Improvements on cross-validation: The .632+ bootstrap method," *J. Amer. Statist. Assoc.*, vol. 92, no. 438, pp. 548–560, 1997.
[39] P. Maji, "f-Information measures for efficient selection of discriminative genes from microarray data," *IEEE Trans. Biomed. Eng.*, vol. 56, no. 4, pp. 1063–1069, Apr. 2009.
[40] S. Paul and P. Maji, "Gene expression and protein-protein interaction data for identification of colon cancer related genes using f-Information measures," *Natural Comput.*, vol. 15, no. 3, pp. 449–463, 2016.

Pradipta Maji (SM'16) received the B.Sc. degree in physics, the M.Sc. degree in electronics science, and the Ph.D. degree in the area of computer science from Jadavpur University, Kolkata, India, in 1998, 2000, and 2005, respectively.

He is currently an Associate Professor in the Machine Intelligence Unit, Indian Statistical Institute, Kolkata, India. His research interests include pattern recognition, machine learning, computational biology and bioinformatics, medical image processing, and so forth. He has published more than 100 papers in international journals and conferences. He is the author of a book published by Wiley-IEEE Computer Society Press and another book published by Springer-Verlag, London.

Dr. Maji has received the 2008 Microsoft Young Faculty Award from Microsoft Research Laboratory India Pvt., the 2009 Young Scientist Award from the National Academy of Sciences, India, the 2011 Young Scientist Award from the Indian National Science Academy, India, and the 2015 Young Faculty Research Fellowship from the Department of Electronics and Information Technology, Government of India. He has been selected as the 2009 Young Associate of the Indian Academy of Sciences, India.

Ankita Mandal received the B.Sc. degree in electronics from West Bengal State University, India, in 2011, and the MCA degree from Jadavpur University, India, in 2014.

She is currently a Research Scholar in the Machine Intelligence Unit, Indian Statistical Institute, Kolkata, India. Her research interests include pattern recognition, machine learning, computational biology and bioinformatics, medical image processing, and so forth. She has published a few papers in international journals and conferences.