FaRoC: Fast and Robust Supervised Canonical Correlation Analysis for Multimodal Omics Data

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

***Abstract*—One of the main problems associated with high dimensional multimodal real life data sets is how to extract relevant and significant features. In this regard, a fast and robust feature extraction algorithm, termed as FaRoC, is proposed, integrating judiciously the merits of canonical correlation analysis (CCA) and rough sets. The proposed method extracts new features sequentially from two multidimensional data sets by maximizing their relevance with respect to class label and significance with respect to already-extracted features. To generate canonical variables sequentially, an analytical formulation is introduced to establish the relation between regularization parameters and CCA. The formulation enables the proposed method to extract required number of correlated features sequentially with lesser computational cost as compared to existing methods. To compute both significance and relevance measures of a feature, the concept of hypercuboid equivalence partition matrix of rough hypercuboid approach is used. It also provides an efficient way to find optimum regularization parameters employed in CCA. The efficacy of the proposed FaRoC algorithm, along with a comparison with other existing methods, is extensively established on several real life data sets.**

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

I. Introduction

IN PRESENT days, there is a scope of getting complementary multiple data corresponding to a given problem or task, and the main challenge is to extract features, which are most relevant, significant, and nonredundant for the given problem. The effective utilization and integration of multiple data sources or multimodal information are becoming an increasingly important problem in many applications. Due to the noisy nature and drastic variation of the acquired signals, unimodal-based pattern recognition and analysis systems usually provide low level of performance, which lead to inaccurate and insufficient pattern representation of the perception of interest. On the other hand, multimodal data contains more information. The integration of multimodal data is expected to provide potentially a more discriminatory and complete description of the intrinsic characteristics of the pattern, which leads to improved system performance than single modality only [1].

The simultaneous analysis of multimodal data sets is an important task in integrative systems biology approach, which gives better understanding of the relationships among different biological functional levels [2]. For example, integration of heterogeneous omics data, namely, transcriptomics, metabolomics, and proteomics, may provide a better understanding of biological systems. The Cancer Genome Atlas (TCGA: http://cancergenome.nih.gov) helps to provide multiple types of data from the same individual. In TCGA, gene and microRNA expression arrays, copy number variation and DNA methylation data, and protein expression array are obtained from most of the tumor samples. By using multiple types of data of unique sample, it is possible to make the linkages between attributes within each type of data. It maximizes the information content and makes a model, which uses all the available data. It is intrinsically more powerful than the models those use only single data type. In this background, there has been an increasing interest in data integration methods in biomedical sciences, both for supervised learning [1], [3]–[5] and unsupervised learning [6]–[9].

Canonical correlation analysis (CCA) [10] provides an efficient way of measuring the linear relationship between two multidimensional data sets. For two multidimensional variables, it finds the best linear transformation to achieve the maximum correlation between them. In many important scientific fields, for example, facial expression recognition, image retrieval, *f*MRI data analysis, and text mining [11]–[13], CCA has been widely applied. In recent years, some variants of CCA, such as generalized CCA [14], kernel CCA [11], sparse CCA [15], and locality preserving CCA [16] have also been developed. The CCA is also popular to integrate different omics data [17]. To map genes or proteins onto the Euclidean space, kernel CCA has been used in [18]. On the other hand, sparse CCA has been used in [19] and [20] to study mutual relation among different types of omics data. Besides of integrating two data sets, CCA can help to analyze gene expression dynamics geometrically [21]. Phylogenetic CCA [22], 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 data sets containing information about a patient's phenotypical and genetic details. It gives those relevant variables or features from

Manuscript received February 2, 2016; revised July 10, 2016; accepted March 11, 2017. Date of publication April 4, 2017; date of current version March 15, 2018. This work was supported by the Department of Electronics and Information Technology, Government of India under Grant PhD-MLA/4(90)/2015-16. This paper was recommended by Associate Editor A. M. Alimi. *(Corresponding author: Pradipta Maji.)*

The authors are with the Biomedical Imaging and Bioinformatics Lab, Machine Intelligence Unit, Indian Statistical Institute, Kolkata 700108, India (e-mail: amandal@isical.ac.in; pmaji@isical.ac.in).

Digital Object Identifier 10.1109/TCYB.2017.2685625

both data types, which are related to each other and provide more insight into the biological experimental hypotheses. Recently, CCA is also used in multiple feature vector fusion, where more than two data sets are used to extract correlated features [23]–[25].

However, CCA suffers from a computational issue due to "large p (number of features) and small n (number of samples)." Let $\mathcal{X}$ and $\mathcal{Y}$ be two multivariate data sets having p and q number of features, respectively, and n is the number of samples in both $\mathcal{X}$ and $\mathcal{Y}$. The features in $\mathcal{X}$ and $\mathcal{Y}$ tend to be highly collinear if $n \ll p$ and $n \ll q$. This leads to ill-conditioned of the covariance matrices of $\mathcal{X}$ and $\mathcal{Y}$, that is, $\mathcal{C}_{xx}$ and $\mathcal{C}_{yy}$. Because of this reason, their inverses are no longer reliable, resulting in an invalid computation of CCA and an unreliable meta-space. The covariance matrices $\mathcal{C}_{xx}$ and $\mathcal{C}_{yy}$ will be invertible if $n \geq p + q + 1$ [26]. However, this condition is usually not possible in the bioinformatics domain, where number of samples n is usually limited. On the other hand, modern technology has enabled very high dimensional data streams to be routinely acquired, which results in very high dimensional feature spaces p and q. To overcome this problem, a regularized version of CCA is introduced in [27]. Regularized CCA (RCCA) [28], [29] is an improved version of CCA. It uses a ridge regression optimization scheme to prevent over-fitting of insufficient training data [30]. It works by adding small positive quantities to the diagonals of $\mathcal{C}_{xx}$ and $\mathcal{C}_{yy}$ to guarantee their invertibility [31]. The RCCA has been successfully used to study gene expressions in liver cells and compare them with concentrations of hepatic fatty acids in mice [29]. Regularized sparse CCA is used in expression quantitative trait loci to detect genetic loci mapped to a disease [32]. However, RCCA is computationally very expensive because of this regularization process. Also, both CCA and RCCA are unsupervised in nature and fail to take complete advantage of available class label information.

To perform the regularization process, supervised RCCA (SRCCA) uses a supervised feature selection algorithm [33]. The available class label information is included in SRCCA to select maximally correlated features. In SRCCA, regularization is done by considering the most discriminatory score of first pair of canonical variables, based on a feature selection method, and then the remaining dimensions are adjusted [33]. One of the important applications of SRCCA in functional genomics is to classify samples, such as to classify cancer versus normal samples or to classify different types or subtypes of cancer, according to the maximally correlated features or biomarkers. SRCCA also helps in developing diagnostic tools for delivering precise, reliable, and interpretable results. With the supervised feature selection results, the cost of biological experiment and decision can be greatly reduced by analyzing only maximally correlated relevant biomarkers. However, existing SRCCA considers only correlation of first pair of canonical variables. 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 for further processing [34]. Recently, a new SRCCA, termed as CCA using maximum relevance-maximum significance criterion and rough sets (CuRSaR) [34], has been proposed, where whole extracted feature set is used to optimize the regularization parameters. However, both existing SRCCA and CuRSaR extract all possible features [$\min(p, q)$], which may not be needed at all. If features are extracted sequentially, then only required number of relevant, significant, and nonredundant features can be extracted. In effect, it will be computationally less expensive.

In integrative omics data analysis, another important problem is uncertainty. This uncertainty may arise from vagueness in response variables of samples and impreciseness in computations. To model and propagate this uncertainty, the theory of rough sets has become successful, which can deal with incompleteness and vagueness [35]. It is proposed for indiscernibility in classification according to some relation and acts as an effective means for dimensionality reduction of discrete valued data [36]. Rough set theory has also been used for analyzing omics data [36]–[43]. Usually, there are continuous valued data in real world applications. In rough set theory, the continuous valued features are divided into several discrete partitions for feature selection. However, the inherent error that exists in discretization process is of major concern in the feature selection. The hypercuboid equivalence partition matrix [44] of rough hypercuboid approach is found to be suitable for feature selection of numerical data. It has been applied successfully for analyzing omics data [34], [45], [46].

In this regard, this paper presents a fast and robust feature extraction algorithm, termed as FaRoC, from two multidimensional data sets. It integrates judiciously the merits of SRCCA and the theory of rough sets. While SRCCA addresses the problem of integrating heterogeneous sources of data, the rough hypercuboid approach of rough sets deals with vagueness in sample categories. The proposed method extracts a new feature by maximizing the relevance with respect to sample categories or class labels and significance with respect to already-extracted features. Both the significance and relevance measures are computed based on the concept of hypercuboid equivalence partition matrix. In the proposed method, the relevance and/or significance do not depend only on the first pair of canonical variables, rather the whole extracted feature set is considered to calculate these measures. A theoretical analysis is presented to establish the relation between CCA and RCCA, which drastically reduces the computational complexity of existing RCCA and helps to extract correlated features sequentially. As the features are extracted sequentially, only required number of significant and relevant features can be generated without generating all possible features. In effect, the proposed method has lower computational cost as compared to existing approaches. The efficacy of the proposed FaRoC algorithm, as well as comparative performance analysis with existing methods, is shown on real life data sets.

The rest of this paper is organized as follows. Section II introduces the basics of CCA, while Section III presents a fast and robust feature extraction algorithm for multimodal data analysis. Experimental results and a comparison among several

existing algorithms are presented in Section IV. Finally, concluding remarks are given in Section V.

II. Basics of CCA, RCCA, and SRCCA

This section presents the fundamental concepts in the theories of CCA, RCCA, and SRCCA.

A. Canonical Correlation Analysis

CCA [10] obtains a linear relationship between two multidimensional variables. The objective of CCA is to extract latent features from two data sets $\mathcal{X} \in \mathbb{R}^{p\times n}$ and $\mathcal{Y} \in \mathbb{R}^{q\times n}$, which are most 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, $w_\mathcal{X} \in \mathbb{R}^p$ and $w_\mathcal{Y} \in \mathbb{R}^q$ such that the empirical correlation between the respective projections onto these weight vectors, that is, between $\mathcal{X}^T w_\mathcal{X}$ and $\mathcal{Y}^T w_\mathcal{Y}$ is maximum. The correlation coefficient $\tilde{\rho}$ is given as follows:

$$\begin{aligned}\tilde{\rho} &= \max_{w_\mathcal{X}, w_\mathcal{Y}} \frac{\mathcal{E}\left[{w_\mathcal{X}}^T \mathcal{X}\mathcal{Y}^T w_\mathcal{Y}\right]}{\sqrt{\mathcal{E}\left[{w_\mathcal{X}}^T \mathcal{X}\mathcal{X}^T w_\mathcal{X}\right]}\sqrt{\mathcal{E}\left[{w_\mathcal{Y}}^T \mathcal{Y}\mathcal{Y}^T w_\mathcal{Y}\right]}} \\ &= \max_{w_\mathcal{X}, w_\mathcal{Y}} \frac{{w_\mathcal{X}}^T C_{\mathcal{X}\mathcal{Y}} w_\mathcal{Y}}{\sqrt{{w_\mathcal{X}}^T C_{\mathcal{X}\mathcal{X}} w_\mathcal{X} {w_\mathcal{Y}}^T C_{\mathcal{Y}\mathcal{Y}} w_\mathcal{Y}}} \end{aligned} \tag{1}$$

where $\mathcal{E}[f]$ denotes empirical expectation of function f, $C_{\mathcal{X}\mathcal{Y}} \in \mathbb{R}^{p\times q}$ is the cross-covariance matrix of $\mathcal{X}$ and $\mathcal{Y}$, which is given as follows:

$$[C_{\mathcal{X}\mathcal{Y}}]_{p\times q} = [\mathcal{X}]_{p\times n}\left[\mathcal{Y}^T\right]_{n\times q} \tag{2}$$

while $C_{\mathcal{X}\mathcal{X}} \in \mathbb{R}^{p\times p}$ and $C_{\mathcal{Y}\mathcal{Y}} \in \mathbb{R}^{q\times q}$ are covariance matrices of $\mathcal{X}$ and $\mathcal{Y}$, respectively, and are as follows:

$$[C_{\mathcal{X}\mathcal{X}}]_{p\times p} = [\mathcal{X}]_{p\times n}\left[\mathcal{X}^T\right]_{n\times p} \tag{3}$$

$$[C_{\mathcal{Y}\mathcal{Y}}]_{q\times q} = [\mathcal{Y}]_{q\times n}\left[\mathcal{Y}^T\right]_{n\times q}. \tag{4}$$

Since $\tilde{\rho}$ is invariant to the scaling of $w_\mathcal{X}$ and $w_\mathcal{Y}$, CCA can be formulated equivalently as

$$\begin{aligned} \max_{w_\mathcal{X}, w_\mathcal{Y}} \quad & {w_\mathcal{X}}^T C_{\mathcal{X}\mathcal{Y}} w_\mathcal{Y} \\ \text{subject to} \quad & {w_\mathcal{X}}^T C_{\mathcal{X}\mathcal{X}} w_\mathcal{X} = 1; \quad \text{and} \quad {w_\mathcal{Y}}^T C_{\mathcal{Y}\mathcal{Y}} w_\mathcal{Y} = 1. \end{aligned} \tag{5}$$

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

$$\Sigma = C_{\mathcal{X}\mathcal{X}}^{-1/2} C_{\mathcal{X}\mathcal{Y}} C_{\mathcal{Y}\mathcal{Y}}^{-1/2}. \tag{6}$$

Suppose $\rho_1 \geq \cdots \geq \rho_i \geq \cdots \geq \rho_p$ be the eigenvalues of $\Sigma\Sigma^T$ and $\xi_{\mathcal{X}1}, \ldots, \xi_{\mathcal{X}i}, \ldots, \xi_{\mathcal{X}p}$ are the orthonormalized eigenvectors corresponding to $\rho_1, \ldots, \rho_i, \ldots, \rho_p$. As nonzero eigenvalues of $\Sigma\Sigma^T$ are same as nonzero eigenvalues of $\Sigma^T\Sigma$ [47], either $\Sigma\Sigma^T$ or $\Sigma^T\Sigma$ is enough to calculate the eigenvectors of them. Furthermore, let say, $p < q$ and $\rho_1 \geq \cdots \geq \rho_i \geq \cdots \geq \rho_p$ are the p largest eigenvalues of $\Sigma^T\Sigma$ with orthonormalized eigenvectors $\xi_{y1}, \ldots, \xi_{yi}, \ldots, \xi_{yp}$. Then, the ith pair of basis vectors are given by

$$w_{\mathcal{X}i} = C_{\mathcal{X}\mathcal{X}}^{-1/2}\xi_{\mathcal{X}i}; \text{ and } w_{\mathcal{Y}i} = C_{\mathcal{Y}\mathcal{Y}}^{-1/2}\xi_{\mathcal{Y}i}. \tag{7}$$

As $\xi_{\mathcal{X}i}$ and $\xi_{\mathcal{Y}i}$ are the ith eigenvectors of $\Sigma\Sigma^T$ and $\Sigma^T\Sigma$, respectively, with eigenvalue ρ_i, the characteristic polynomials of $\Sigma\Sigma^T$ and $\Sigma^T\Sigma$ can be written as

$$\begin{aligned} \Sigma\Sigma^T\xi_{\mathcal{X}i} &= \rho_i\xi_{\mathcal{X}i} \\ \Rightarrow \left(C_{\mathcal{X}\mathcal{X}}^{-1/2} C_{\mathcal{X}\mathcal{Y}} C_{\mathcal{Y}\mathcal{Y}}^{-1/2}\right)\left(C_{\mathcal{X}\mathcal{X}}^{-1/2} C_{\mathcal{X}\mathcal{Y}} C_{\mathcal{Y}\mathcal{Y}}^{-1/2}\right)^T\xi_{\mathcal{X}i} &= \rho_i\xi_{\mathcal{X}i} \\ \Rightarrow C_{\mathcal{X}\mathcal{X}}^{-1/2} C_{\mathcal{X}\mathcal{Y}} C_{\mathcal{Y}\mathcal{Y}}^{-1/2} C_{\mathcal{Y}\mathcal{Y}}^{-1/2} {C_{\mathcal{X}\mathcal{Y}}}^T C_{\mathcal{X}\mathcal{X}}^{-1/2}\xi_{\mathcal{X}i} &= \rho_i\xi_{\mathcal{X}i} \\ \Rightarrow C_{\mathcal{X}\mathcal{X}}^{-1/2} C_{\mathcal{X}\mathcal{Y}} C_{\mathcal{Y}\mathcal{Y}}^{-1} C_{\mathcal{Y}\mathcal{X}} C_{\mathcal{X}\mathcal{X}}^{-1/2}\xi_{\mathcal{X}i} &= \rho_i\xi_{\mathcal{X}i} \\ \Rightarrow C_{\mathcal{X}\mathcal{X}}^{-1/2} C_{\mathcal{X}\mathcal{X}}^{-1/2} C_{\mathcal{X}\mathcal{Y}} C_{\mathcal{Y}\mathcal{Y}}^{-1} C_{\mathcal{Y}\mathcal{X}} C_{\mathcal{X}\mathcal{X}}^{-1/2}\xi_{\mathcal{X}i} &= \rho_i C_{\mathcal{X}\mathcal{X}}^{-1/2}\xi_{\mathcal{X}i} \\ \Rightarrow C_{\mathcal{X}\mathcal{X}}^{-1} C_{\mathcal{X}\mathcal{Y}} C_{\mathcal{Y}\mathcal{Y}}^{-1} C_{\mathcal{Y}\mathcal{X}} w_{\mathcal{X}i} &= \rho_i w_{\mathcal{X}i} \end{aligned} \tag{8}$$

and

$$\begin{aligned} \Sigma^T\Sigma\xi_{\mathcal{Y}i} &= \rho_i\xi_{\mathcal{Y}i} \\ \Rightarrow C_{\mathcal{Y}\mathcal{Y}}^{-1} C_{\mathcal{Y}\mathcal{X}} C_{\mathcal{X}\mathcal{X}}^{-1} C_{\mathcal{X}\mathcal{Y}} w_{\mathcal{Y}i} &= \rho_i w_{\mathcal{Y}i}. \end{aligned} \tag{9}$$

From (8) and (9), it can be seen that basis vectors $w_{\mathcal{X}i}$ and $w_{\mathcal{Y}i}$ are the eigenvectors of matrix $\mathcal{H}$ and $\tilde{\mathcal{H}}$, respectively, with eigenvalue ρ_i, where

$$\mathcal{H} = C_{\mathcal{X}\mathcal{X}}^{-1} C_{\mathcal{X}\mathcal{Y}} C_{\mathcal{Y}\mathcal{Y}}^{-1} C_{\mathcal{Y}\mathcal{X}}; \quad \text{and} \quad \tilde{\mathcal{H}} = C_{\mathcal{Y}\mathcal{Y}}^{-1} C_{\mathcal{Y}\mathcal{X}} C_{\mathcal{X}\mathcal{X}}^{-1} C_{\mathcal{X}\mathcal{Y}}. \tag{10}$$

The ith pair of canonical variables $\{\mathcal{U}_i, \mathcal{V}_i\}$ is as follows:

$$\mathcal{U}_i = {w_{\mathcal{X}i}}^T\mathcal{X}; \quad \text{and} \quad \mathcal{V}_i = {w_{\mathcal{Y}i}}^T\mathcal{Y}. \tag{11}$$

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 ith basis vectors having unit variance and data set. It maximizes the correlation among all possible linear combinations and is uncorrelated with the previous $(i-1)$ canonical variable pairs. From (11), the ith feature $\mathcal{A}_i$ is extracted as follows:

$$\mathcal{A}_i = \mathcal{U}_i + \mathcal{V}_i \tag{12}$$

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

B. Regularized Canonical Correlation Analysis

Functional genomics experiments generally give rise to very complicated data, which are often plagued with noise. RCCA [28], [29] is used to correct these noises in $\mathcal{X}$ and $\mathcal{Y}$. Let us assume that $\mathcal{X}$ and $\mathcal{Y}$ are contaminated with Gaussian, independent and identically distributed noise $\mathcal{N}_\mathcal{X} \in \mathbb{R}^{p\times n}$ and $\mathcal{N}_\mathcal{Y} \in \mathbb{R}^{q\times n}$. As these noises are Gaussian, independent and identically distributed, all possible combinations of the covariances of the p and q rows of $\mathcal{N}_\mathcal{X}$ and $\mathcal{N}_\mathcal{Y}$, respectively, will be 0 except the covariance of a particular row vector with itself. Let the variances of each row of $\mathcal{N}_\mathcal{X}$ and $\mathcal{N}_\mathcal{Y}$ be $\mathfrak{r}_\mathcal{X}$ and $\mathfrak{r}_\mathcal{Y}$, respectively, which are known as regularization parameters. The cross-covariance matrix $C_{\mathcal{X}\mathcal{Y}}$ of $\mathcal{X}$ and $\mathcal{Y}$ will not be affected. But, the matrices $C_{\mathcal{X}\mathcal{X}}$ and $C_{\mathcal{Y}\mathcal{Y}}$ become $[C_{\mathcal{X}\mathcal{X}} + \mathfrak{r}_\mathcal{X} I]$ and $[C_{\mathcal{Y}\mathcal{Y}} + \mathfrak{r}_\mathcal{Y} I]$, respectively, where I is the identity matrix. So, (10) becomes

$$\mathcal{H} = \left[C_{\mathcal{X}\mathcal{X}} + \mathfrak{r}_\mathcal{X} I\right]^{-1} C_{\mathcal{X}\mathcal{Y}}\left[C_{\mathcal{Y}\mathcal{Y}} + \mathfrak{r}_\mathcal{Y} I\right]^{-1} C_{\mathcal{Y}\mathcal{X}} \tag{13}$$

and

$$\tilde{\mathcal{H}} = \left[C_{yy} + \mathfrak{r}_y I\right]^{-1} C_{yx} \left[C_{xx} + \mathfrak{r}_x I\right]^{-1} C_{xy}. \tag{14}$$

In RCCA, the regularization parameters are varied in a certain range $\mathfrak{r}_{\min} \leq \mathfrak{r}_x, \mathfrak{r}_y \leq \mathfrak{r}_{\max}$ and chosen by a grid search optimization technique [48]. 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

$$\max_{\mathfrak{r}_x, \mathfrak{r}_y} \frac{w_x{}^T C_{xy} w_y}{\sqrt{w_x{}^T \left(C_{xx} + \mathfrak{r}_x I\right) w_x w_y{}^T (C_{yy} + \mathfrak{r}_y I) w_y}}. \tag{15}$$

C. Supervised Regularized Canonical Correlation Analysis

Both CCA and RCCA are unsupervised in nature. They do not incorporate the information of class label or sample category even if it is present in the given data sets. To overcome this limitation of both CCA and RCCA, Golugula *et al.* [33] introduced the concept of SRCCA, which is a supervised version of RCCA. Similar to RCCA, SRCCA chooses the optimal regularization parameters $\mathfrak{r}_x$ and $\mathfrak{r}_y$ using grid search optimization by a feature selection method based on either *t*-test, Wilks's lambda test, or Wilcoxon rank sum test. The optimal regularization parameters are obtained by maximizing the discriminatory score of the feature corresponding to first pair of canonical variables, and then the remaining dimensions are extracted for the optimal parameters.

III. Proposed Method

This section presents a fast and robust feature extraction algorithm, termed as FaRoC, integrating judiciously the information of two multidimensional data sets. Some important analytical formulations are reported next prior to describing the proposed method.

A. Relation Between CCA and RCCA

The spectral decomposition [49] can be used to calculate $[C_{xx}+\mathfrak{r}_x I]^{-1}$ and $[C_{yy}+\mathfrak{r}_y I]^{-1}$ for the computation of $\mathcal{H}$ matrix of (13). 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 [49]:

$$\left[C_{xx} + \mathfrak{r}_x I\right] = \Psi_x \Lambda_x \Psi_x^T = \sum_{i=1}^{p} \lambda_i \psi_i \psi_i^T \tag{16}$$

where the ith element λ_i of diagonal matrix Λ_x denotes the ith eigenvalue of the matrix $[C_{xx} + \mathfrak{r}_x I]$. The ith column of matrix Ψ_x represents the orthonormalized eigenvector ψ_i corresponding to eigenvalue λ_i, $\forall i \in \{1, 2, \ldots, p\}$, and

$$\Psi_x \Psi_x^T = \Psi_x^T \Psi_x = I. \tag{17}$$

The computation of the inverse of matrix $[C_{xx} + \mathfrak{r}_x I]$ is performed as follows [50]:

$$\left[C_{xx} + \mathfrak{r}_x I\right]^{-1} = \Psi_x \Lambda_x^{-1} \Psi_x^T = \sum_{i=1}^{p} \frac{1}{\lambda_i} \psi_i \psi_i^T. \tag{18}$$

In general, the regularized parameters $\mathfrak{r}_x$ and $\mathfrak{r}_y$ of both RCCA and SRCCA are varied within a specified range $[\mathfrak{r}_{\min}, \mathfrak{r}_{\max}]$, where $\mathfrak{r}_{\min} \leq \mathfrak{r}_x, \mathfrak{r}_y \leq \mathfrak{r}_{\max}$. Let us assume that these regularization parameters follow an arithmetic progression. Initially, $\mathfrak{r}_x = \mathfrak{r}_{\min}$ and $\mathfrak{r}_y = \mathfrak{r}_{\min}$ and at final step $\mathfrak{r}_x + (\mathfrak{t}_x - 1)\mathfrak{d}_x = \mathfrak{r}_{\max}$ and $\mathfrak{r}_y + (\mathfrak{t}_y - 1)\mathfrak{d}_y = \mathfrak{r}_{\max}$. Here, $\mathfrak{d}_x$ and $\mathfrak{d}_y$ denote the common differences for regularization parameters $\mathfrak{r}_x$ and $\mathfrak{r}_y$, respectively, while the parameters $\mathfrak{t}_x$ and $\mathfrak{t}_y$ are the number of possible values of $\mathfrak{r}_x$ and $\mathfrak{r}_y$. It is clearly seen that the diagonal elements of the covariance matrices are only changed by adding regularization parameters. If a regularization parameter is added on the diagonal elements of the covariance matrix, the eigenvalues are changed, but the eigenvectors remain same [34]. In other words, if Λ_{x_i} and Λ_{y_j}, respectively, denote the $(i-1)$th and $(j-1)$th diagonal matrices, where diagonal elements are the eigenvalues of $[C_{xx} + (\mathfrak{r}_x + (i-1)\mathfrak{d}_x) I]$ and $[C_{yy} + (\mathfrak{r}_y + (j-1)\mathfrak{d}_y) I]$, then

$$\Lambda_{x_i} = \Lambda_x + (i-1)\mathfrak{d}_x I \tag{19}$$

$$\Lambda_{y_j} = \Lambda_y + (j-1)\mathfrak{d}_y I \tag{20}$$

where $\Lambda_x = \Lambda_{x_1}$, $\Lambda_y = \Lambda_{y_1}$, and

$$[C_{xx} + (\mathfrak{r}_x + i\mathfrak{d}_x) I]\Psi_x = \Psi_x(\Lambda_x + i\mathfrak{d}_x I) \tag{21}$$

$$[C_{yy} + (\mathfrak{r}_y + j\mathfrak{d}_y) I]\Psi_y = \Psi_y(\Lambda_y + j\mathfrak{d}_y I). \tag{22}$$

Based on the above analysis, it can be shown that if the regularization parameters $\mathfrak{r}_x$ and $\mathfrak{r}_y$ follow an arithmetic progression, the matrix $\mathcal{H}$ of (13) and the matrix $\tilde{\mathcal{H}}$ of (14) become

$$\mathcal{H}_{ij} = \left[C_{xx} + (\mathfrak{r}_x + (i-1)\mathfrak{d}_x) I\right]^{-1} \times C_{xy}\left[C_{yy} + (\mathfrak{r}_y + (j-1)\mathfrak{d}_y) I\right]^{-1} C_{yx} \tag{23}$$

and

$$\tilde{\mathcal{H}}_{ij} = \left[C_{yy} + (\mathfrak{r}_y + (j-1)\mathfrak{d}_y) I\right]^{-1} \times C_{yx}\left[C_{xx} + (\mathfrak{r}_x + (i-1)\mathfrak{d}_x) I\right]^{-1} C_{xy} \tag{24}$$

where $\forall i \in \{1, 2, \ldots, \mathfrak{t}_x\}$ and $\forall j \in \{1, 2, \ldots, \mathfrak{t}_y\}$. Combining (18) and (21)–(23), we get

$$\mathcal{H}_{ij} = \Psi_x [\Lambda_x + (i-1)\mathfrak{d}_x I]^{-1} \Psi_x^T C_{xy} \Psi_y \left[\Lambda_y + (j-1)\mathfrak{d}_y I\right]^{-1} \Psi_y^T C_{yx}. \tag{25}$$

Similarly, combining (18), (21), (22), and (24), we get

$$\tilde{\mathcal{H}}_{ij} = \Psi_y \left[\Lambda_y + (j-1)\mathfrak{d}_y I\right]^{-1} \Psi_y^T C_{yx} \Psi_x \left[\Lambda_x + (i-1)\mathfrak{d}_x I\right]^{-1} \Psi_x^T C_{xy}. \tag{26}$$

Here, $[\Lambda_x + (i-1)\mathfrak{d}_x I]$ is a nonsingular diagonal matrix, which is obtained by adding two diagonal matrices, namely, Λ_x and $[(i-1)\mathfrak{d}_x I]$. The diagonal elements of Λ_x represent the eigenvalues of matrix C_{xx}, while that of $[(i-1)\mathfrak{d}_x I]$ are $(i-1)\mathfrak{d}_x$. As Λ_x and $[\Lambda_x + (i-1)\mathfrak{d}_x I]$ are nonsingular matrices, and $[(i-1)\mathfrak{d}_x I]$ has rank p for $i > 1$, the inverse of matrix $[\Lambda_x + (i-1)\mathfrak{d}_x I]$ can be calculated using the inverse of matrix Λ_x [51]. As matrix $[(i-1)\mathfrak{d}_x I]$ has rank p, the matrix $[\Lambda_x + (i-1)\mathfrak{d}_x I]$ can be written as

$$\mathcal{G}_{p+1} = \Lambda_x + (i-1)\mathfrak{d}_x I = \Lambda_x + [(i-1)\mathfrak{d}_x I]_1 + [(i-1)\mathfrak{d}_x I]_2 + \cdots + [(i-1)\mathfrak{d}_x I]_p \tag{27}$$

where each $[(i-1)d_{\mathcal{X}}I]_r$, $\forall r = 1, 2, \ldots, p$, has rank 1. So, the inverse of $\mathcal{G}_{p+1}$ can be expressed as follows:

$$\mathcal{G}_{p+1}^{-1} = \mathcal{G}_p^{-1} + g_p \mathcal{G}_p^{-1}[(i-1)d_{\mathcal{X}}I]_p \mathcal{G}_p^{-1}$$
$$\Rightarrow \mathcal{G}_{p+1}^{-1} = \Lambda_{\mathcal{X}}^{-1} + \sum_{r=1}^{p} g_r \mathcal{G}_r^{-1}[(i-1)d_{\mathcal{X}}I]_r \mathcal{G}_r^{-1} \quad (28)$$

considering $\mathcal{G}_1 = \Lambda_{\mathcal{X}}$, where

$$g_r = \frac{1}{1 + \text{trace}\left(\mathcal{G}_r^{-1}[(i-1)d_{\mathcal{X}}I]_r\right)}. \quad (29)$$

Similarly, the matrix $[\Lambda_y + (j-1)d_y I]$ can be written as

$$\tilde{\mathcal{G}}_{q+1} = \Lambda_y + (j-1)d_y I = \Lambda_y + [(j-1)d_y I]_1 + [(j-1)d_y I]_2 + \cdots + [(j-1)d_y I]_q \quad (30)$$

where the matrix $[(j-1)d_y I]$ has rank q for $j > 1$ and each $[(j-1)d_y I]_s$, $\forall s = 1, 2, \ldots, q$, has rank 1. So, the inverse of $\tilde{\mathcal{G}}_{q+1}$ can be expressed, considering $\tilde{\mathcal{G}}_1 = \Lambda_y$, as follows:

$$\tilde{\mathcal{G}}_{q+1}^{-1} = \Lambda_y^{-1} + \sum_{s=1}^{q} \tilde{g}_s \tilde{\mathcal{G}}_s^{-1}[(j-1)d_y I]_s \tilde{\mathcal{G}}_s^{-1} \quad (31)$$

where

$$\tilde{g}_s = \frac{1}{1 + \text{trace}\left(\tilde{\mathcal{G}}_s^{-1}[(j-1)d_y I]_s\right)}. \quad (32)$$

Hence, using (28) and (31), the matrix $\mathcal{H}$ of (25) becomes

$$\mathcal{H}_{ij} = \Psi_{\mathcal{X}} \left(\Lambda_{\mathcal{X}}^{-1} + \sum_{r=1}^{p} g_r \mathcal{G}_r^{-1}[(i-1)d_{\mathcal{X}}I]_r \mathcal{G}_r^{-1} \right) \Psi_{\mathcal{X}}^T C_{\mathcal{X}y} \Psi_y \times \left(\Lambda_y^{-1} + \sum_{s=1}^{q} \tilde{g}_s \tilde{\mathcal{G}}_s^{-1}[(j-1)d_y I]_s \tilde{\mathcal{G}}_s^{-1} \right) \Psi_y^T C_{y\mathcal{X}}$$
$$\Rightarrow \mathcal{H}_{ij} = \Psi_{\mathcal{X}} \Lambda_{\mathcal{X}}^{-1} \Psi_{\mathcal{X}}^T C_{\mathcal{X}y} \Psi_y \Lambda_y^{-1} \Psi_y^T C_{y\mathcal{X}} + \mathcal{B}_{ij} = \mathcal{H}_{11} + \mathcal{B}_{ij} \quad (33)$$

where

$$\mathcal{B}_{ij} = \Theta_i \Psi_y \Lambda_y^{-1} \Psi_y^T C_{y\mathcal{X}} + \Psi_{\mathcal{X}} \Lambda_{\mathcal{X}}^{-1} \Psi_{\mathcal{X}}^T C_{\mathcal{X}y} \Phi_j + \Theta_i \Phi_j$$
$$\Rightarrow \mathcal{B}_{ij} = \Theta_i C_{yy}^{-1} C_{y\mathcal{X}} + C_{\mathcal{X}\mathcal{X}}^{-1} C_{\mathcal{X}y} \Phi_j + \Theta_i \Phi_j \quad (34)$$

$$\Theta_i = \Psi_{\mathcal{X}} \sum_{r=1}^{p} g_r \mathcal{G}_r^{-1}[(i-1)d_{\mathcal{X}}I]_r \mathcal{G}_r^{-1} \Psi_{\mathcal{X}}^T C_{\mathcal{X}y} \quad (35)$$

and

$$\Phi_j = \Psi_y \sum_{s=1}^{q} \tilde{g}_s \tilde{\mathcal{G}}_s^{-1}[(j-1)d_y I]_s \tilde{\mathcal{G}}_s^{-1} \Psi_y^T C_{y\mathcal{X}}. \quad (36)$$

Similarly, using (28) and (31), the matrix $\tilde{\mathcal{H}}$ of (26) becomes

$$\tilde{\mathcal{H}}_{ij} = \Psi_y \left(\Lambda_y^{-1} + \sum_{s=1}^{q} \tilde{g}_s \tilde{\mathcal{G}}_s^{-1}[(j-1)d_y I]_s \tilde{\mathcal{G}}_s^{-1} \right) \Psi_y^T C_{y\mathcal{X}} \Psi_{\mathcal{X}} \left(\Lambda_{\mathcal{X}}^{-1} + \sum_{r=1}^{p} g_r \mathcal{G}_r^{-1}[(i-1)d_{\mathcal{X}}I]_r \mathcal{G}_r^{-1} \right) \Psi_{\mathcal{X}}^T C_{\mathcal{X}y}$$
$$\Rightarrow \tilde{\mathcal{H}}_{ij} = \Psi_y \Lambda_y^{-1} \Psi_y^T C_{y\mathcal{X}} \Psi_{\mathcal{X}} \Lambda_{\mathcal{X}}^{-1} \Psi_{\mathcal{X}}^T C_{\mathcal{X}y} + \tilde{\mathcal{B}}_{ij} = \tilde{\mathcal{H}}_{11} + \tilde{\mathcal{B}}_{ij} \quad (37)$$

where

$$\tilde{\mathcal{B}}_{ij} = \Phi_j \Psi_{\mathcal{X}} \Lambda_{\mathcal{X}}^{-1} \Psi_{\mathcal{X}}^T C_{\mathcal{X}y} + \Psi_y \Lambda_y^{-1} \Psi_y^T C_{y\mathcal{X}} \Theta_i + \Phi_j \Theta_i$$
$$\Rightarrow \tilde{\mathcal{B}}_{ij} = \Phi_j C_{\mathcal{X}\mathcal{X}}^{-1} C_{\mathcal{X}y} + C_{yy}^{-1} C_{y\mathcal{X}} \Theta_i + \Phi_j \Theta_i. \quad (38)$$

From (33) and (37), it is clear that if eigenvalues and eigenvectors of $C_{\mathcal{X}\mathcal{X}}$ and C_{yy} are calculated to compute $\mathcal{H}_{11}$ and $\tilde{\mathcal{H}}_{11}$ matrices for initial values of $\mathfrak{r}_{\mathcal{X}}$ and $\mathfrak{r}_y$, there is no need to compute eigenvalues and eigenvectors for computing $\mathcal{H}_{ij}$ and $\tilde{\mathcal{H}}_{ij}$ at other values of $\mathfrak{r}_{\mathcal{X}}$ and $\mathfrak{r}_y$, as initial eigenvalues and eigenvectors can be used to compute different $\mathcal{H}_{ij}$ and $\tilde{\mathcal{H}}_{ij}$ matrices. Also, if the minimum value of $\mathfrak{r}_{\mathcal{X}}$ and $\mathfrak{r}_y$ is set to 0, then eigenvalues and eigenvectors of CCA can be used to compute different $\mathcal{H}_{ij}$ and $\tilde{\mathcal{H}}_{ij}$ matrices of RCCA corresponding to different values of regularization parameters.

B. Sequential Generation of Canonical Variables

From (8) and (9), it is evident that the nonzero eigenvalues of $\Sigma\Sigma^T$, $\Sigma^T\Sigma$, $\mathcal{H}$, and $\tilde{\mathcal{H}}$ are same [47]. So, either $\mathcal{H}$ or $\tilde{\mathcal{H}}$ is computed using (33) or (37), respectively, corresponding to a pair of regularization parameters $\mathfrak{r}_{\mathcal{X}}$ and $\mathfrak{r}_y$ depending on whether $p \leq q$ or $p > q$. Let us assume that $\mathcal{H}$ has kth eigenvalue ρ_k and corresponding eigenvector is $w_{\mathcal{X}k}$. So

$$\mathcal{H} w_{\mathcal{X}k} = \rho_k w_{\mathcal{X}k}$$
$$\Rightarrow C_{\mathcal{X}\mathcal{X}}^{-1} C_{\mathcal{X}y} C_{yy}^{-1} C_{y\mathcal{X}} w_{\mathcal{X}k} = \rho_k w_{\mathcal{X}k}$$
$$\Rightarrow C_{yy}^{-1} C_{y\mathcal{X}} C_{\mathcal{X}\mathcal{X}}^{-1} C_{\mathcal{X}y} C_{yy}^{-1} C_{y\mathcal{X}} w_{\mathcal{X}k} = \rho_k C_{yy}^{-1} C_{y\mathcal{X}} w_{\mathcal{X}k}$$
$$\Rightarrow \tilde{\mathcal{H}} w_{yk} = \rho_k w_{yk}; \quad \text{where} \quad w_{yk} = C_{yy}^{-1} C_{y\mathcal{X}} w_{\mathcal{X}k}. \quad (39)$$

So, the kth eigenvector w_{yk} of $\tilde{\mathcal{H}}$ is proportional to $C_{yy}^{-1} C_{y\mathcal{X}}$ and can be obtained from the kth eigenvector $w_{\mathcal{X}k}$ of $\mathcal{H}$ using (39). So, from (39), it is also clear that either $\mathcal{H}$ or $\tilde{\mathcal{H}}$ is enough to calculate the eigenvectors of $\mathcal{H}$ and $\tilde{\mathcal{H}}$. Assuming $\mathcal{K} = \min(p, q)$, $\mathcal{K}$ eigenvalue-eigenvector pairs can be calculated using Jacobi method [52]. Then, $\mathcal{K}$ pairs of basis vectors and $\mathcal{K}$ pairs of canonical variables are computed using (8) or (9) and (11), respectively. Finally, $\mathcal{K}$ features can be extracted using (12). The computational complexity of Jacobi method to compute $\mathcal{K}$ eigenvalue-eigenvector pairs is $\mathcal{O}(\mathcal{K}^3)$.

However, the value of $\mathcal{K}$ is large for real life high dimensional multimodal data analysis. So, a small fraction, among the huge amount of extracted features, is effective to perform a certain task. Furthermore, a small subset of extracted features is advisable to develop tools for delivering interpretable, reliable, and precise results. Hence, the goal of multimodal data analysis is to identify a reduced set of most relevant extracted features. This is referred to as feature selection, and an important problem in machine learning. So, instead of generating all $\mathcal{K}$ eigenvalue-eigenvector pairs using Jacobi method, if each eigenvalue-eigenvector pair of $\mathcal{H}$ is generated sequentially, the quality of each extracted feature can be evaluated, and finally, $\mathcal{D}$ features can be extracted for multimodal data analysis, where $\mathcal{D} \ll \mathcal{K}$. In the proposed method, each eigenvalue-eigenvector pair of $\mathcal{H}$ is calculated sequentially by using power method [52]. The kth eigenvalue-eigenvector pair can be calculated with the help of first eigenvalue-eigenvector

pair as explained below. Following analysis establishes that there is a direct relation between kth and $(k+1)$th eigenvalue-eigenvector pairs, and using this relation, all correlated features can be extracted sequentially. Let us assume that ρ_k and w_{xk} be the kth eigenvalue and corresponding eigenvector, respectively, of $\mathcal{H}$ matrix. So,

$$\begin{aligned}\mathcal{H}w_{xk} &= \rho_k w_{xk} \\ &\Rightarrow \mathcal{H}w_{xk}w_{xk}^T = \rho_k w_{xk}w_{xk}^T \\ &\Rightarrow \mathcal{H} - \mathcal{H}w_{xk}w_{xk}^T = \mathcal{H} - \rho_k w_{xk}w_{xk}^T \\ &\Rightarrow \left(\mathcal{H} - \mathcal{H}w_{xk}w_{xk}^T\right)w_{x(k+1)} \\ &= \left(\mathcal{H} - \rho_k w_{xk}w_{xk}^T\right)w_{x(k+1)} \\ &\Rightarrow \mathcal{H}w_{x(k+1)} - \mathcal{H}w_{xk}w_{xk}^T w_{x(k+1)} \\ &= \left(\mathcal{H} - \rho_k w_{xk}w_{xk}^T\right)w_{x(k+1)} \\ &\Rightarrow \mathcal{H}w_{x(k+1)} \\ &= \left(\mathcal{H} - \rho_k w_{xk}w_{xk}^T\right)w_{x(k+1)} \end{aligned} \tag{40}$$

$$\tag{41}$$

where $w_{x(k+1)}$ is the $(k+1)$th eigenvector of $\mathcal{H}$ corresponding to the eigenvalue $\rho_{(k+1)}$, that is,

$$\mathcal{H}w_{x(k+1)} = \rho_{(k+1)}w_{x(k+1)}. \tag{42}$$

Hence, from (41) and (42), we get

$$(\mathcal{H} - \rho_k w_{xk}w_{xk}^T)w_{x(k+1)} = \rho_{(k+1)}w_{x(k+1)}. \tag{43}$$

Hence, from (43), it is proved that the $(k+1)$th eigenvalue-eigenvector pair $\{\rho_{(k+1)}, w_{x(k+1)}\}$ of the matrix $\mathcal{H}$ is same as first eigenvalue-eigenvector pair of the matrix $(\mathcal{H}-\rho_k w_{xk}w_{xk}^T)$. For calculating $(k+1)$th eigenvalue-eigenvector pair, the matrices $\mathcal{H}$ and $\tilde{\mathcal{H}}$ can be calculated, based on Deflation method [53], as follows:

$$\mathcal{H}(k+1) = \mathcal{H}(k) - \rho_k w_{xk}w_{xk}^T = \mathcal{H}(1) - \sum_{l=1}^{k}\rho_l w_{xl}w_{xl}^T \tag{44}$$

$$\tilde{\mathcal{H}}(k+1) = \tilde{\mathcal{H}}(k) - \rho_k w_{yk}w_{yk}^T = \tilde{\mathcal{H}}(1) - \sum_{l=1}^{k}\rho_l w_{yl}w_{yl}^T. \tag{45}$$

Therefore, $\rho_{(k+1)}$ and $w_{x(k+1)}$ can be calculated with the help of previously calculated eigenvalue-eigenvector pairs, that is, ρ_l and w_{xl}, $\forall l = 1, 2, \ldots, k$. Hence, using (44), each eigenvalue-eigenvector pair of matrix $\mathcal{H}$ can be calculated sequentially. So, for RCCA with (i,j)th regularization parameters of $\mathfrak{r}_x$ and $\mathfrak{r}_y$, to compute $(k+1)$th basis eigenvector, the matrices $\mathcal{H}_{ij}$ and $\tilde{\mathcal{H}}_{ij}$ can be calculated by using (33), (44) and (37), (45) as follows:

$$\begin{aligned}\mathcal{H}_{ij}(k+1) &= \mathcal{H}_{ij}(1) - \sum_{l=1}^{k}\rho_{l_{ij}}w_{xl_{ij}}w_{xl_{ij}}^T \\ \Rightarrow \mathcal{H}_{ij}(k+1) &= \mathcal{H}_{11} + \mathcal{B}_{ij} - \sum_{l=1}^{k}\rho_{l_{ij}}w_{xl_{ij}}w_{xl_{ij}}^T \end{aligned} \tag{46}$$

and

$$\begin{aligned}\tilde{\mathcal{H}}_{ij}(k+1) &= \tilde{\mathcal{H}}_{ij}(1) - \sum_{l=1}^{k}\rho_{l_{ij}}w_{yl_{ij}}w_{yl_{ij}}^T \\ \Rightarrow \tilde{\mathcal{H}}_{ij}(k+1) &= \tilde{\mathcal{H}}_{11} + \tilde{\mathcal{B}}_{ij} - \sum_{l=1}^{k}\rho_{l_{ij}}w_{yl_{ij}}w_{yl_{ij}}^T \end{aligned} \tag{47}$$

where $\forall k \in \{1, 2, \ldots, \mathcal{K}\}$, $\mathcal{K} = \min(p, q)$, $\forall i \in \{1, 2, \ldots, \mathfrak{t}_x\}$, and $\forall j \in \{1, 2, \ldots, \mathfrak{t}_y\}$.

C. Relevance and Significance for Regularization

One of the major concerns in high dimensional multimodal real data analysis is how to extract relevant and significant features. In general, a huge number of irrelevant and insignificant features may be present in the extracted feature set, which may degrade the classification accuracy by reducing the useful information. So, the features those are highly relevant with respect to sample categories or class labels and have high significance in the feature set should be considered in the extracted feature subset. The class labels of the unknown samples are expected to be predicted correctly by such features. Therefore, a measure is needed by which the quality of a set of features can be evaluated. In the current work, significant and relevant features are extracted from two multidimensional data sets using hypercuboid equivalence partition matrix.

Let $\mathcal{X} \in \mathbb{R}^{p\times n}$ and $\mathcal{Y} \in \mathbb{R}^{q\times n}$ be two multidimensional data sets with p and q variables or attributes, respectively, and n samples. Let us assume that each attribute 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 $[\mathfrak{r}_{\min}, \mathfrak{r}_{\max}]$, where $\mathfrak{r}_{\min} \le \mathfrak{r}_x, \mathfrak{r}_y \le \mathfrak{r}_{\max}$. 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 $\gamma_{\mathcal{A}_k}(\mathbb{D})$ is 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 $\mathcal{D} \le \min(p, q)$. The change in joint relevance or dependency when a feature is discarded from the set of features, is a measure of the significance of the feature. To what extent a feature contributes for computing the dependency on class labels can be computed by the significance of the feature. The significance of the feature $\mathcal{A}_k$ with respect to the feature set $\{\mathcal{A}_k, \mathcal{A}_l\}$ is given by

$$\sigma_{\{\mathcal{A}_k,\mathcal{A}_l\}}(\mathbb{D}, \mathcal{A}_k) = \gamma_{\{\mathcal{A}_k,\mathcal{A}_l\}}(\mathbb{D}) - \gamma_{\mathcal{A}_l}(\mathbb{D}). \tag{48}$$

So, the higher change in dependency indicates that the feature $\mathcal{A}_k$ is more significant. If significance is 0, then the feature is dispensable.

Hence, the problem of extracting a relevant and significant feature set $\mathbb{S}$ from all possible combinations of regularization parameters $\mathfrak{r}_x$ and $\mathfrak{r}_y$ is equivalent to maximize the average relevance of all extracted features as well as to maximize the average significance among them. The following greedy algorithm is used to solve the above problem.

1) Calculate the cross-covariance matrix $\mathcal{C}_{xy} \in \mathbb{R}^{p\times q}$ of $\mathcal{X}$ and $\mathcal{Y}$ using (2).
2) Calculate covariance matrices $\mathcal{C}_{xx} \in \mathbb{R}^{p\times p}$ and $\mathcal{C}_{yy} \in \mathbb{R}^{q\times q}$ of $\mathcal{X}$ and $\mathcal{Y}$ using (3) and (4), respectively.
3) Calculate eigenvalues $\Lambda_x \in \mathbb{R}^p$ and $\Lambda_y \in \mathbb{R}^q$ of $\mathcal{C}_{xx}$ and $\mathcal{C}_{yy}$, respectively, along with corresponding eigenvectors Ψ_x and Ψ_y using Jacobi method.

4) If $p \leq q$, calculate $\mathcal{H}_{11}$ using (25), otherwise calculate $\tilde{\mathcal{H}}_{11}$ using (26).
5) Initialize $\mathbb{S} \leftarrow \emptyset$ and $\mathcal{k} = 1$.
6) Repeat the following three steps until $\mathcal{k} \leq \mathcal{D}$.
   a) Take $\mathbb{C} \leftarrow \emptyset$.
   b) Repeat the following seven steps for all (i, j)th regularization parameters of $\mathfrak{r}_\mathcal{X}$ and $\mathfrak{r}_y$, where $\forall i \in \{1, 2, \ldots, \mathfrak{t}_\mathcal{X}\}$ and $\forall j \in \{1, 2, \ldots, \mathfrak{t}_y\}$.
      For $p \leq q$ (respectively, $q > p$).
      i) If $\mathcal{k} = 1$, calculate $\mathcal{H}_{ij}(\mathcal{k})$ using (33) [respectively, $\tilde{\mathcal{H}}_{ij}(\mathcal{k})$ using (37)], otherwise using (46) [respectively, using (47)].
      ii) Calculate largest eigenvalue $\rho_{\mathcal{k}_{ij}}$ and eigenvector $w_{\mathcal{X}\mathcal{k}_{ij}}$ (respectively, $w_{y\mathcal{k}_{ij}}$) of matrix $\mathcal{H}_{ij}(\mathcal{k})$ [respectively, $\tilde{\mathcal{H}}_{ij}(\mathcal{k})$] using Power method and (39), where $w_{\mathcal{X}\mathcal{k}_{ij}}$ and $w_{y\mathcal{k}_{ij}}$ are the $\mathcal{k}$th basis vectors.
      iii) Calculate the $\mathcal{k}$th pair of canonical variables $\{\mathcal{U}_{\mathcal{k}_{ij}}, \mathcal{V}_{\mathcal{k}_{ij}}\}$ using (11).
      iv) Compute the $\mathcal{k}$th extracted feature $\mathcal{A}_{\mathcal{k}_{ij}}$ corresponding to (i, j)th pair of regularization parameters using (12).
      v) Calculate the relevance $\gamma_{\mathcal{A}_{\mathcal{k}_{ij}}}(\mathbb{D})$ of $\mathcal{A}_{\mathcal{k}_{ij}}$.
      vi) Calculate the significance $\sigma_{\{\mathcal{A}_{\mathcal{k}_{ij}}, \mathcal{A}_\ell\}}(\mathbb{D}, \mathcal{A}_{\mathcal{k}_{ij}})$ of $\mathcal{A}_{\mathcal{k}_{ij}}$ with respect to each $\mathcal{A}_\ell$ of the already selected features of $\mathbb{S}$ using (48).
      vii) Add $\mathcal{A}_{\mathcal{k}_{ij}}$ to $\mathbb{C}$ if its significance is nonzero with respect to all of the selected features of $\mathbb{S}$. In effect, $\mathbb{C} = \mathbb{C} \cup \mathcal{A}_{\mathcal{k}_{ij}}$.
   c) If $\mathbb{C} \neq \emptyset$, select a feature as $\mathcal{k}$th feature $\mathcal{A}_\mathcal{k}$ from all the features of $\mathbb{C}$, which maximizes the following condition:

$$\begin{cases} \gamma_{\mathcal{A}_{\mathcal{k}_{ij}}}(\mathbb{D}) & \text{if } \mathcal{k} = 1 \\ \gamma_{\mathcal{A}_{\mathcal{k}_{ij}}}(\mathbb{D}) + \frac{1}{\mathcal{k}-1} \sum\limits_{\mathcal{A}_\ell \in \mathbb{S}} \sigma_{\{\mathcal{A}_{\mathcal{k}_{ij}}, \mathcal{A}_\ell\}}\left(\mathbb{D}, \mathcal{A}_{\mathcal{k}_{ij}}\right) & \text{otherwise.} \end{cases} \tag{49}$$

      As a result of that, $\mathbb{S} = \mathbb{S} \cup \mathcal{A}_\mathcal{k}$ and $\mathcal{k} = \mathcal{k} + 1$.
7) Stop.

The concept of hypercuboid equivalence partition matrix [44] of rough hypercuboid approach, presented next, is used to compute both relevance and significance of an extracted feature. The regularization parameters are optimized through computing these two measures.

D. Computation of Relevance and Significance

Generally, an m-dimensional hypercuboid or hyperrectangle is defined in the m-dimensional Euclidean space, where the space is defined by the m variables measured for each sample or object. In geometry, a hypercuboid or hyperrectangle is the generalization of a rectangle for higher dimensions, formally defined as the Cartesian product of orthogonal intervals. A d-dimensional hypercuboid with d attributes as its dimensions is defined as the Cartesian product of d orthogonal intervals. It encloses a region in the d-dimensional space, where each dimension corresponds to a certain attribute. The value domain of each dimension is the value range or interval that corresponds to a particular class. For all hypercuboids, any two objects belong to a same class hypercuboid are said to be indiscernible with respect to that particular class.

Let $\mathbb{U} = \{O_1, \ldots, O_j, \ldots, O_n\}$ be the set of n samples or objects with condition attribute or feature set $\mathbb{C} = \{\mathcal{A}_1, \ldots, \mathcal{A}_\mathcal{k}, \ldots, \mathcal{A}_\mathfrak{t}\}$, where $\mathfrak{t} \leq (\mathfrak{t}_\mathcal{X} \times \mathfrak{t}_y)$ is the total number of extracted candidate features, at a particular iteration, having nonzero significance values with respect to the already selected features of $\mathbb{S}$, $\mathfrak{t}_\mathcal{X}$ and $\mathfrak{t}_y$ denote the number of possible values of regularization parameters $\mathfrak{r}_\mathcal{X}$ and $\mathfrak{r}_y$, respectively. Let $\mathbb{D}$ be the class label or decision attribute set. If $\mathbb{U}/\mathbb{D} = \{\beta_1, \ldots, \beta_i, \ldots, \beta_c\}$ denotes c equivalence classes or granules of the universe $\mathbb{U}$ created by the equivalence relation induced from $\mathbb{D}$, then c information granules of $\mathbb{U}$ can also be created by the equivalence relation induced from each condition attribute $\mathcal{A}_\mathcal{k} \in \mathbb{C}$. If $\mathbb{U}/\mathcal{A}_\mathcal{k} = \{\delta_1, \ldots, \delta_i, \ldots, \delta_c\}$ denotes c equivalence classes or information granules of $\mathbb{U}$ induced by the condition attribute $\mathcal{A}_\mathcal{k}$ and n is the number of objects in $\mathbb{U}$, then c-partitions of $\mathbb{U}$ are the sets of (cn) values $\{\hbar_{ij}(\mathcal{A}_\mathcal{k})\}$, which are arrayed as a matrix $\mathbb{H}(\mathcal{A}_\mathcal{k}) = [\hbar_{ij}(\mathcal{A}_\mathcal{k})]_{c\times n}$. The matrix $\mathbb{H}(\mathcal{A}_\mathcal{k})$ is termed as hypercuboid equivalence partition matrix of the condition attribute $\mathcal{A}_\mathcal{k}$ [44], where

$$\hbar_{ij}(\mathcal{A}_\mathcal{k}) = \begin{cases} 1 & \text{if } \mathcal{L}_i \leq O_j(\mathcal{A}_\mathcal{k}) \leq \mathcal{U}_i \\ 0 & \text{otherwise} \end{cases} \tag{50}$$

represents the membership of object O_j in the ith equivalence partition or class β_i satisfying the following two conditions:

$$1 \leq \sum_{j=1}^{n} \hbar_{ij}(\mathcal{A}_\mathcal{k}) \leq n, \forall i; \quad 1 \leq \sum_{i=1}^{c} \hbar_{ij}(\mathcal{A}_\mathcal{k}) \leq c, \forall j. \tag{51}$$

Here, $[\mathcal{L}_i, \mathcal{U}_i]$ represents the interval of ith class β_i according to the class labels $\mathbb{D}$. The interval $[\mathcal{L}_i, \mathcal{U}_i]$ is spanned by the objects with class β_i with respect to the condition attribute $\mathcal{A}_\mathcal{k}$. In other words, the value of each object $O_j \in \beta_i$ with respect to $\mathcal{A}_\mathcal{k}$ falls within $[\mathcal{L}_i, \mathcal{U}_i]$. A $c \times n$ hypercuboid equivalence partition matrix $\mathbb{H}(\mathcal{A}_\mathcal{k})$ represents the c-hypercuboid equivalence partitions of the universe generated by an equivalence relation. Each row of this matrix represents a hypercuboid equivalence class or partition. The ith hypercuboid partition is represented as follows [44]:

$$\beta_i = \left\{\hbar_{i1}(\mathcal{A}_\mathcal{k})/O_1 + \hbar_{i2}(\mathcal{A}_\mathcal{k})/O_2 + \cdots + \hbar_{in}(\mathcal{A}_\mathcal{k})/O_n\right\}. \tag{52}$$

However, every two intervals or hypercuboids may intersect with each other. These intersections form the implicit hypercuboids, which encompass objects those are misclassified. The degree of dependency of a condition attribute or a subset of attributes on decision attribute is estimated based on the cardinality of implicit hypercuboids. The misclassified objects belonging to implicit hypercuboids are identified using the confusion vector, which is defined based on hypercuboid equivalence partition matrix as follows [44]:

$$\mathbb{V}(\mathcal{A}_\mathcal{k}) = \left[v_1(\mathcal{A}_\mathcal{k}), v_2(\mathcal{A}_\mathcal{k}), \ldots, v_n(\mathcal{A}_\mathcal{k})\right] \tag{53}$$

where

$$v_j(\mathcal{A}_\mathcal{k}) = \min\left\{1, \sum_{i=1}^{c} \hbar_{ij}(\mathcal{A}_\mathcal{k}) - 1\right\}. \tag{54}$$

In other words, $v_j(\mathcal{A}_k) = 1$ if the jth object O_j belongs to the implicit hypercuboid, which represents the boundary region of more than one classes. On the other hand, if the object O_j is encompassed by the lower approximation of any class, then $v_j(\mathcal{A}_k) = 0$ and the object O_j does not belong to the lower or upper approximations of any other classes. Hence, the confusion vector and hypercuboid equivalence partition matrix corresponding to feature $\mathcal{A}_k$ can be used for defining upper and lower approximations of the class β_i. Let $\beta_i \subseteq \mathbb{U}$. The information of the attribute $\mathcal{A}_k$ can be used to approximate β_i, by constructing $\mathcal{R}$-lower approximation and $\mathcal{R}$-upper approximation of β_i:

$$\underline{\mathcal{R}}(\beta_i) = \left\{O_j |\ h_{ij}(\mathcal{A}_k) = 1 \text{ and } v_j(\mathcal{A}_k) = 0\right\} \tag{55}$$

$$\overline{\mathcal{R}}(\beta_i) = \left\{O_j |\ h_{ij}(\mathcal{A}_k) = 1\right\} \tag{56}$$

where the attribute $\mathcal{A}_k$ induces equivalence relation $\mathcal{R}$. Hence, the cardinality of lower approximation of class β_i is computed as follows:

$$\left|\underline{\mathcal{R}}(\beta_i)\right| = \sum_{j=1}^{n} h_{ij}(\mathcal{A}_k)\left[1 - v_j(\mathcal{A}_k)\right]. \tag{57}$$

Based on the definition of lower approximation of rough sets, the positive region of decision attribute set $\mathbb{D}$ is defined as

$$\text{POS}_{\mathcal{R}}(\mathbb{D}) = \bigcup_{\beta_i \in \mathbb{U}/\mathbb{D}} \underline{\mathcal{R}}(\beta_i). \tag{58}$$

The positive region, $\text{POS}_{\mathcal{R}}(\mathbb{D})$, contains all objects of $\mathbb{U}$ that can be classified to classes of $\mathbb{U}/\mathbb{D}$ using the knowledge in attribute $\mathcal{A}_k$. Combining (50), (53), and (58), the cardinality of positive regions of decision attribute $\mathbb{D}$, in terms of hypercuboid equivalence partition matrix and confusion vector of condition attribute $\mathcal{A}_k$, is given by

$$|\text{POS}_{\mathcal{R}}(\mathbb{D})| = \sum_{i=1}^{c}\sum_{j=1}^{n} h_{ij}(\mathcal{A}_k)\left[1 - v_j(\mathcal{A}_k)\right]. \tag{59}$$

Hence, the dependency between condition attribute $\mathcal{A}_k$ and decision attribute $\mathbb{D}$ is as follows:

$$\gamma_{\mathcal{A}_k}(\mathbb{D}) = \frac{1}{n}\sum_{i=1}^{c}\sum_{j=1}^{n} h_{ij}(\mathcal{A}_k)\left[1 - v_j(\mathcal{A}_k)\right] \tag{60}$$

that is

$$\gamma_{\mathcal{A}_k}(\mathbb{D}) = 1 - \frac{1}{n}\sum_{j=1}^{n} v_j(\mathcal{A}_k) \tag{61}$$

where $\gamma_{\mathcal{A}_k}(\mathbb{D}) \in [0, 1]$. If $\mathbb{D}$ depends totally on $\mathcal{A}_k$, then $\gamma_{\mathcal{A}_k}(\mathbb{D}) = 1$; if $\mathbb{D}$ depends partially on $\mathcal{A}_k$, then $\gamma_{\mathcal{A}_k}(\mathbb{D}) \in (0, 1)$; and if $\mathbb{D}$ does not depend on $\mathcal{A}_k$, then $\gamma_{\mathcal{A}_k}(\mathbb{D}) = 0$.

The relevance of a feature $\mathcal{A}_k$ with respect to the class label or decision attribute $\mathbb{D}$ is computed using (61), while the joint relevance $\gamma_{\{\mathcal{A}_k,\mathcal{A}_l\}}(\mathbb{D})$ is to be computed to calculate the significance of the feature $\mathcal{A}_k$ with respect to the set $\{\mathcal{A}_k, \mathcal{A}_l\}$ using (48). The joint relevance depends on the $c \times n$ hypercuboid equivalence partition matrix corresponding to the set $\{\mathcal{A}_k, \mathcal{A}_l\}$, which is computed from two $c \times n$ equivalence partition matrices $\mathbb{H}(\mathcal{A}_k)$ and $\mathbb{H}(\mathcal{A}_l)$ as follows:

$$\mathbb{H}(\{\mathcal{A}_k, \mathcal{A}_l\}) = \mathbb{H}(\mathcal{A}_k) \cap \mathbb{H}(\mathcal{A}_l) \tag{62}$$

where

$$h_{ij}(\{\mathcal{A}_k, \mathcal{A}_l\}) = h_{ij}(\mathcal{A}_k) \times h_{ij}(\mathcal{A}_l). \tag{63}$$

E. Computational Complexity

Let $\mathcal{X}$ and $\mathcal{Y}$ be the two datasets with n samples and c classes, where p and q represent the number of features in $\mathcal{X}$ and $\mathcal{Y}$, respectively. Let us assume that the regularization parameters $\mathfrak{r}_\mathcal{X}$ and $\mathfrak{r}_y$ have $\mathfrak{t}_\mathcal{X}$ and $\mathfrak{t}_y$ possible values. Let $\mathcal{M} = \max(p, q)$ and $\mathcal{K} = \min(p, q)$, where the number of extracted features $\mathcal{D} \ll \mathcal{K}$. The computational complexity to calculate cross-covariance matrix $\mathcal{C}_{\mathcal{X}y}$ is $\mathcal{O}(\mathcal{K}\mathcal{M}n)$, whereas the total time complexity to compute covariance matrices $\mathcal{C}_{\mathcal{X}\mathcal{X}}$ and $\mathcal{C}_{yy}$ is $\mathcal{O}(\mathcal{K}^2n + \mathcal{M}^2n)$. In step 3, the eigenvalues $\Lambda_\mathcal{X}$ and Λ_y, along with corresponding eigenvectors $\Psi_\mathcal{X}$ 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 $(\mathcal{O}(\mathcal{K}\mathcal{M}n + \mathcal{K}^2n + \mathcal{M}^2n + \mathcal{K}^3 + \mathcal{M}^3) =)\mathcal{O}(\mathcal{M}^3)$. The total time complexity to compute $\mathcal{C}_{\mathcal{X}\mathcal{X}}^{-1}$ and $\mathcal{C}_{yy}^{-1}$ is $\mathcal{O}(\mathcal{K}^3 + \mathcal{M}^3)$. So, step 4, for computing the matrix $\mathcal{H}_{11}$, has computational complexity $(\mathcal{O}(\mathcal{K}^3 + \mathcal{K}^2\mathcal{M} + \mathcal{K}\mathcal{M}^2 + \mathcal{M}^3) =)\mathcal{O}(\mathcal{M}^3)$. Step 5 has constant time complexity, which is $\mathcal{O}(1)$.

There is a loop in step 6, which is executed $\mathcal{D}$ times. The first step of this loop has constant complexity of $\mathcal{O}(1)$ and the next step has another loop, which is executed $(\mathfrak{t}_\mathcal{X} \times \mathfrak{t}_y)$ times. The computational complexity to calculate $\mathcal{B}_{ij}$ or $\tilde{\mathcal{B}}_{ij}$ is $(\mathcal{O}(\mathcal{K}^2 + \mathcal{M}^2 + \mathcal{K}^2\mathcal{M} + \mathcal{K}\mathcal{M}^2) =)\mathcal{O}(\mathcal{K}\mathcal{M}^2)$. Hence, the total complexity of step 6) b) i) is $(\mathcal{O}(\mathcal{K}\mathcal{M}^2 + \mathcal{K}^2) =)\mathcal{O}(\mathcal{K}\mathcal{M}^2)$. The next step has $\mathcal{O}(\mathcal{K}^2)$ time complexity to calculate the eigenvalue and corresponding eigenvector (which is a basis vector) using Power method. On the other hand, another basis vector can be calculated with time complexity $\mathcal{O}(\mathcal{K}\mathcal{M}^2 + \mathcal{K}\mathcal{M})$. So, step 6) b) ii) has total complexity $(\mathcal{O}(\mathcal{K}^2 + \mathcal{K}\mathcal{M}^2 + \mathcal{K}\mathcal{M}) =)$ $\mathcal{O}(\mathcal{K}\mathcal{M}^2)$. The total time complexity for computing canonical variables $\mathcal{U}$ and $\mathcal{V}$ in step 6) b) iii) is $\mathcal{O}(\mathcal{K}n + \mathcal{M}n)$. The computational complexity to extract a feature $\mathcal{A}$ is $\mathcal{O}(n)$. The time complexity to compute both relevance and significance of a feature is same, which is $\mathcal{O}(cn)$. Hence, the total complexity to execute the loop $(\mathfrak{t}_\mathcal{X} \times \mathfrak{t}_y)$ times is $(\mathcal{O}(\mathfrak{t}_\mathcal{X}\mathfrak{t}_y(\mathcal{K}\mathcal{M}^2 + \mathcal{K}\mathcal{M}^2 + \mathcal{K}n + \mathcal{M}n + n + cn)) =)\mathcal{O}(\mathfrak{t}_\mathcal{X}\mathfrak{t}_y\mathcal{K}\mathcal{M}^2)$. The selection of a feature from $(\mathfrak{t}_\mathcal{X} \times \mathfrak{t}_y)$ candidate features by maximizing relevance and significance, which is carried out in step 6) c), has complexity $\mathcal{O}(\mathfrak{t}_\mathcal{X}\mathfrak{t}_y)$. Hence, the total complexity to execute the loop $\mathcal{D}$ times is $(\mathcal{O}(\mathcal{D}(\mathfrak{t}_\mathcal{X}\mathfrak{t}_y\mathcal{K}\mathcal{M}^2 + \mathfrak{t}_\mathcal{X}\mathfrak{t}_y)) =)$ $\mathcal{O}(\mathcal{D}\mathfrak{t}_\mathcal{X}\mathfrak{t}_y\mathcal{K}\mathcal{M}^2)$. Hence, the overall computational complexity of the proposed algorithm is $(\mathcal{O}(\mathcal{M}^3 + \mathcal{M}^3 + \mathcal{D}\mathfrak{t}_\mathcal{X}\mathfrak{t}_y\mathcal{K}\mathcal{M}^2) =)$ $\mathcal{O}(\mathcal{M}^2(\mathcal{M} + \mathcal{D}\mathfrak{t}_\mathcal{X}\mathfrak{t}_y\mathcal{K}))$.

IV. Experimental Results and Discussion

The performance of the proposed feature extraction algorithm, termed as FaRoC, is extensively studied and compared with that of some existing CCA-based algorithms. The algorithms compared are CCA, RCCA, several variants of SRCCA using t-test (SRCCA_{TT}) [33], Wilcoxon rank sum test (SRCCA_{WR}) [33], Wilks's lambda test (SRCCA_{WL}) [33], mutual information (SRCCA_{MI}), rough hypercuboid (SRCCA_{RH}) and CuRSaR [34]. The performance

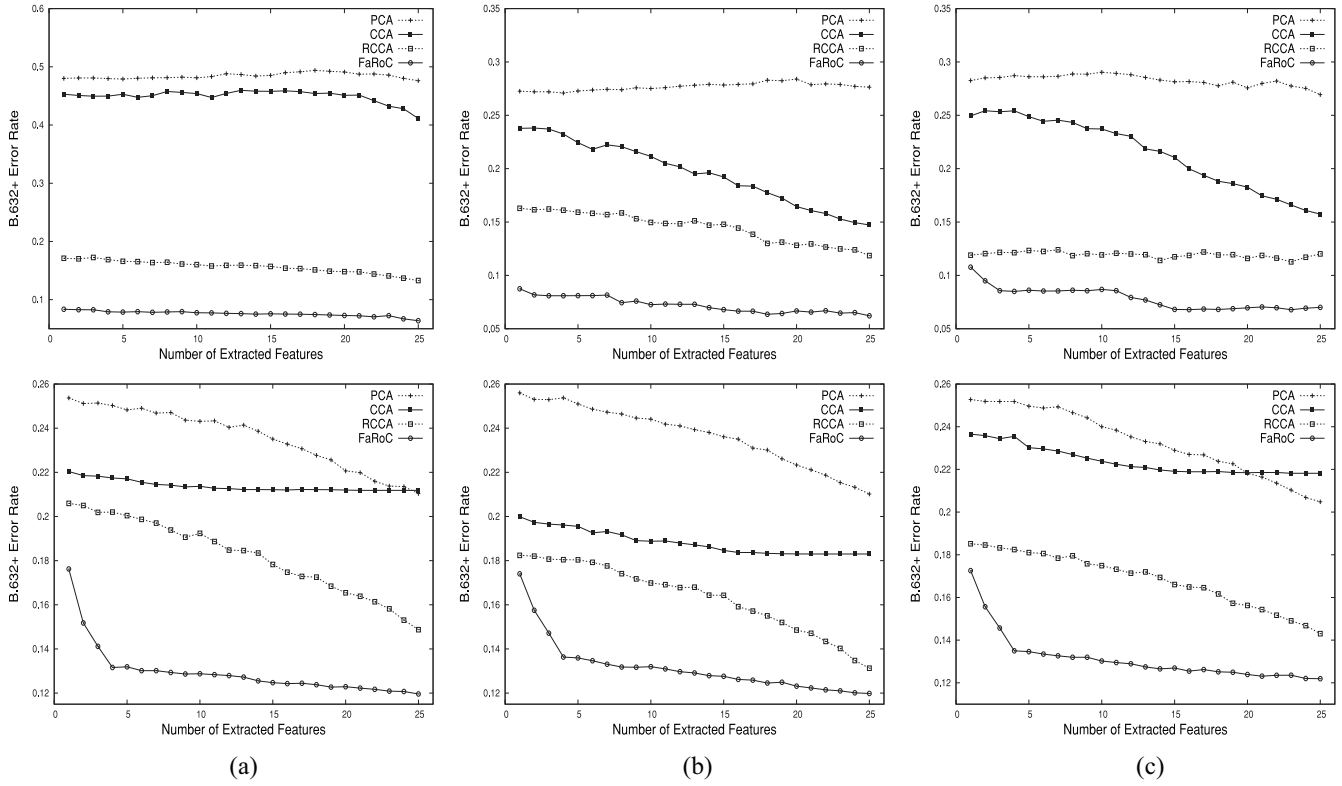

Fig. 1. Comparative performance analysis between PCA, CCA, RCCA, and FaRoC (top: BRCA and bottom: OV). (a) Gene-DNA methylation. (b) Gene-protein. (c) Protein-DNA methylation.

of rough hypercuboid approach is also compared with that of mutual information in the proposed feature extraction framework. The $B.632+$ bootstrap approach [54] is used in order to minimize the biasedness and variability of obtained results. The support vector machine (SVM) [55] with linear kernels is used to compute this error. The source code of the proposed FaRoC algorithm, written in C language, is available at www.isical.ac.in/∼bibl/results/faroc/faroc.html.

A. Description of Data Sets

Two multimodal data sets, 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 data sets are downloaded from TCGA. The BRCA data set contains a total number of 204 breast invasive carcinoma samples, classified into two classes: 1) 189 samples of infiltrating ductal carcinoma and 2) 15 samples of infiltrating lobular carcinoma. On the other hand, OV data set consists of 379 ovarian serous cystadenocarcinoma samples, grouped into two categories: 1) 51 samples of grade 2 and 2) 328 samples of grade 3 ovarian serous cystadenocarcinoma. Both data sets 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. Effectiveness of Proposed FaRoC Algorithm

In this section, the performance of the FaRoC algorithm is presented on both BRCA and OV data sets considering three pairs of modalities, namely, gene-protein, gene-DNA methylation, and protein-DNA methylation. The performance of the proposed method is compared with that of PCA, CCA, and RCCA. Results are reported in Fig. 1 with respect to $B.632+$ error rate of the SVM considering 25 extracted features, while Table I compares the minimum error rates and average cosine distance obtained using different methods. The value of each regularization parameter varies from 0.0 to 1.0 with a difference 0.1.

From the results reported in Fig. 1, it is clearly observed that the $B.632+$ error rate for the proposed method decreases as the number of extracted features increases. Also, the $B.632+$ error rate of the FaRoC algorithm is significantly lower as compared to existing PCA, CCA, and RCCA, irrespective of the data sets, pair of modalities, and number of extracted features. The results reported in Table I also show that the minimum error rate obtained using the proposed FaRoC algorithm is significantly lower compared to both CCA and RCCA, irrespective of data and modalities; however, the average cosine distance among extracted features of the FaRoC is lower. Also, the performance of the PCA is significantly poor compared to CCA, RCCA, and FaRoC due to the drastic variation and noisy nature of different modalities. The significantly better performance of the proposed algorithm is achieved due to the fact that the FaRoC algorithm extracts features sequentially by maximizing both significance and relevance of the features. Both of these 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,

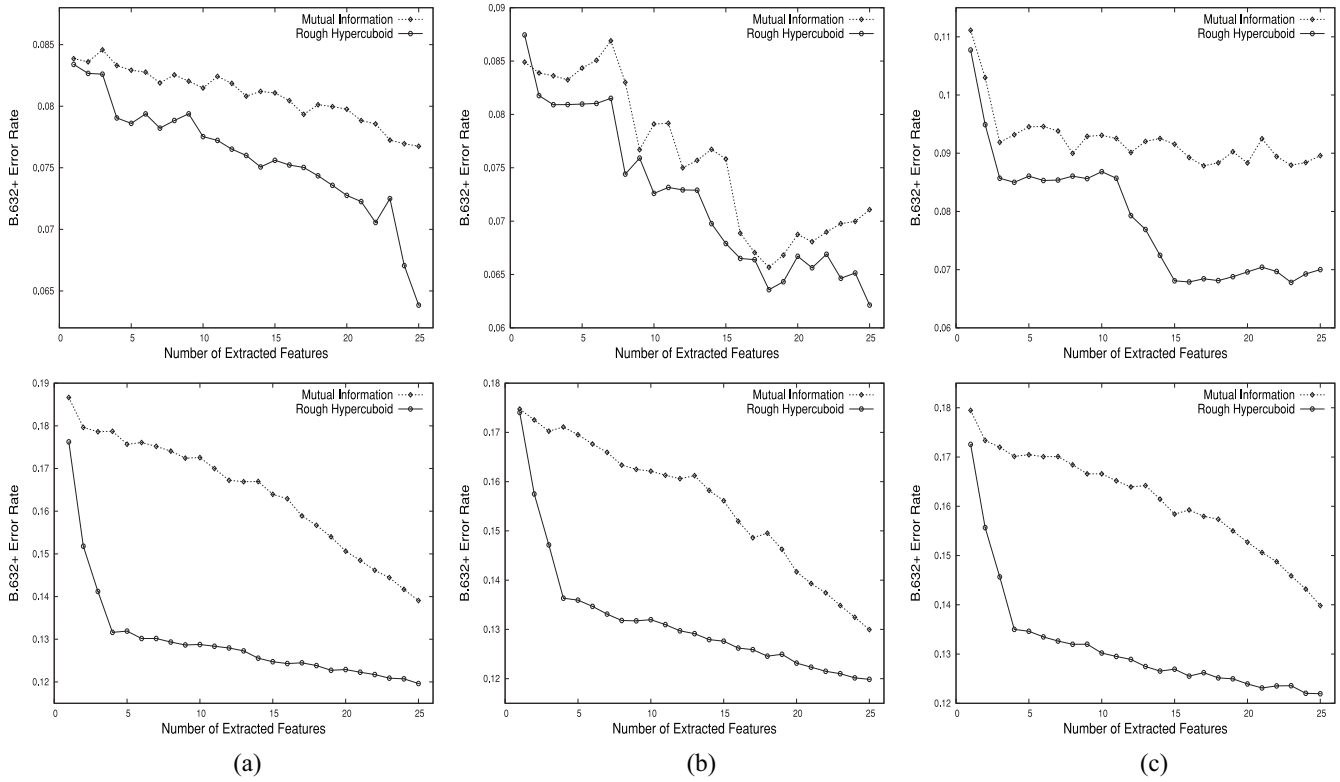

Fig. 2. Comparative performance analysis between mutual information and rough hypercuboid in proposed method (top: BRCA and bottom: OV). (a) Gene-DNA methylation. (b) Gene-protein. (c) Protein-DNA methylation.

the proposed algorithm is able to extract more relevant and significant features from a pair of modalities.

C. Importance of Rough Hypercuboid Approach

In the proposed FaRoC algorithm, both relevance and significance measures of an extracted feature are computed based on the concept of hypercuboid equivalence partition matrix. The relevance of an extracted feature with respect to decision attribute set or class labels is calculated as per (61), while significance of a feature is calculated using (48) with respect to the already-extracted features. In this regard, it should be noted that other measures like mutual information can also be employed for computing both significance and relevance of a feature. Fig. 2 establishes the importance of rough hypercuboid approach over mutual information considering two data sets and three pairs of modalities. Subsequent discussions analyze the results with respect to $B.632+$ error rate of the SVM.

All the results reported in Fig. 2 confirm that the performance of hypercuboid equivalence partition matrix is significantly better than that of mutual information, irrespective of the data sets, number of extracted features, and pairs of modalities. Also, the minimum error rates obtained using mutual information for BRCA data are 0.065679, 0.087837, and 0.076747, respectively, for gene-protein, protein-DNA methylation, and gene-DNA methylation, respectively, while that of rough hypercuboid approach are 0.062137, 0.067812, and 0.063843. Similarly, for OV data set, the minimum error rates of mutual information are 0.129967, 0.139822, and 0.139059, respectively, while that of rough hypercuboid approach are 0.119867, 0.121938, and 0.119630. The significantly better performance of the rough hypercuboid-based proposed approach is obtained due to the fact that the quality of an extracted feature set, in rough hypercuboid approach, is evaluated by the hypercuboid equivalence partition matrix that makes use of supervised information of sample categories in granulation process. Also, it provides an efficient way to calculate relevance and significance in approximation spaces. The proposed FaRoC algorithm, in effect, is able to generate a reduced set of significant and relevant features from multimodal omics data sets.

D. Importance of Sequential Feature Generation

The proposed FaRoC algorithm extracts $\mathcal{D}$ features sequentially from two multidimensional data sets, based on their individual relevance with respect to class label and significance with respect to the already-extracted features. However, $\mathcal{D}$ features can be extracted simultaneously by maximum relevance-maximum significance criterion as done in CuRSaR [34]. In order to establish the importance of sequential feature generation of the proposed FaRoC algorithm over simultaneous feature generation by CuRSaR, extensive experimental results are reported in Fig. 3. The results reported in Fig. 3 establish the fact that the FaRoC outperforms CuRSaR in almost all the cases, irrespective of data sets, pair of modalities, and number of extracted features. Also, the minimum error obtained by the FaRoC algorithm, as reported in Table I, is significantly lower as compared to CuRSaR. The better performance of the FaRoC algorithm over CuRSaR is achieved due to the fact that the FaRoC considers different pairs of regularization parameters for different features, while the CuRSaR extracts

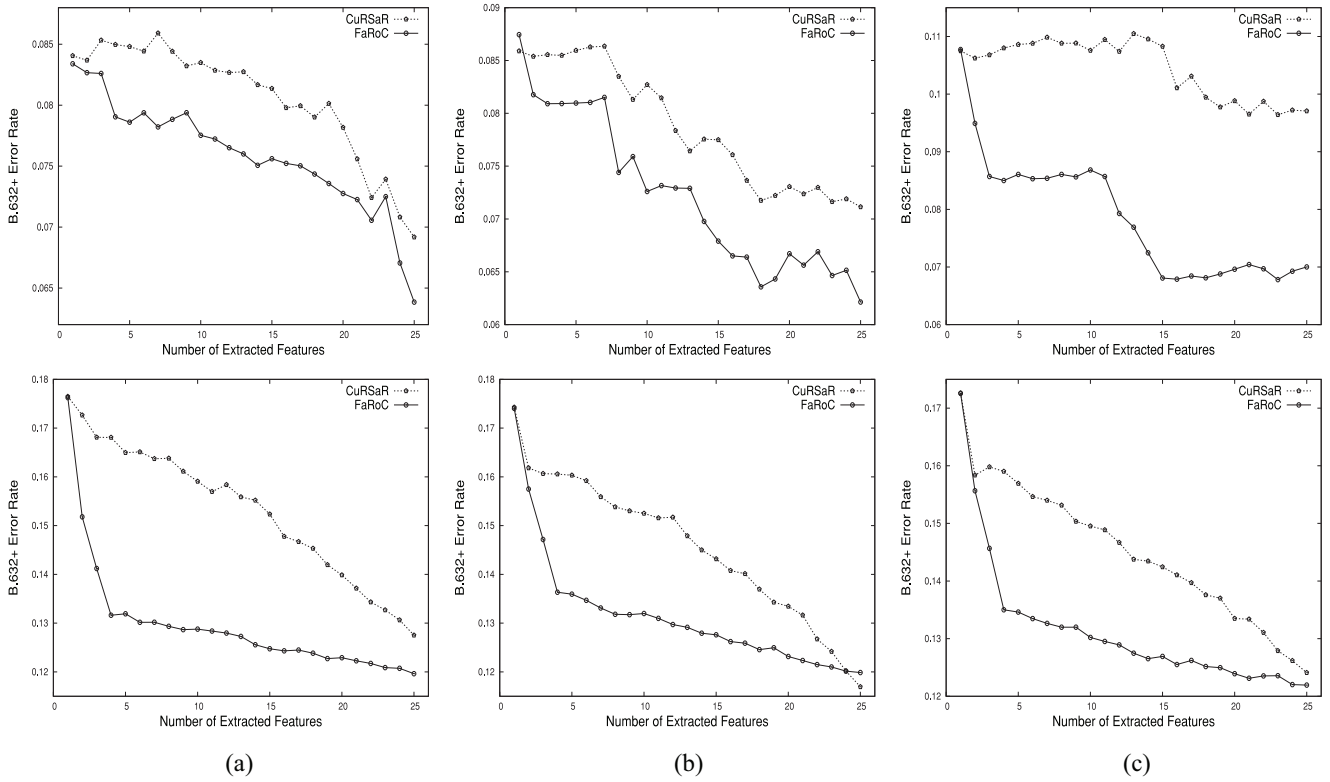

Fig. 3. Comparative performance analysis between CuRSaR and FaRoC algorithms (top: BRCA and bottom: OV). (a) Gene-DNA methylation. (b) Gene-protein. (c) Protein-DNA methylation.

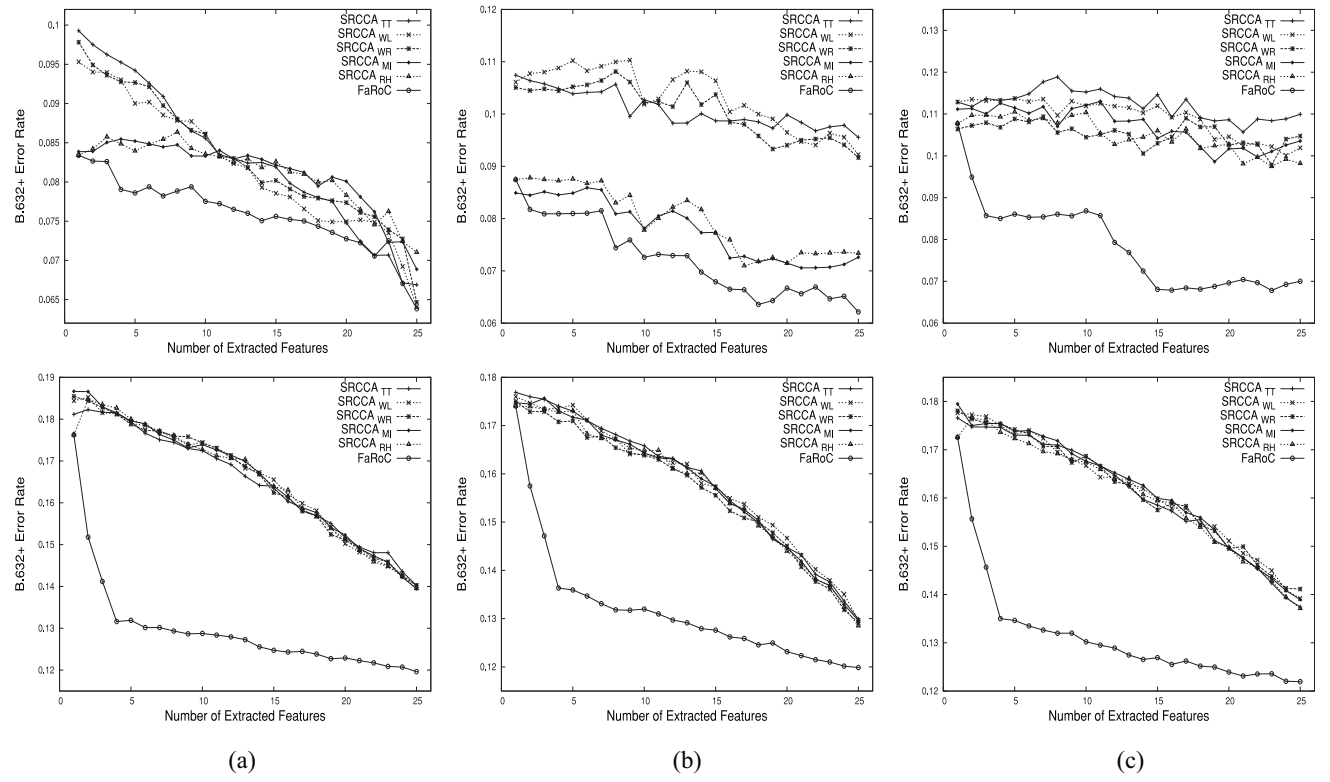

Fig. 4. Comparative performance analysis between different SRCCA methods and proposed FaRoC algorithm (top: BRCA and bottom: OV). (a) Gene-DNA methylation. (b) Gene-protein. (c) Protein-DNA methylation.

set of features for a fixed pair of parameters. In effect, the extracted features are more relevant and significant for the FaRoC than the CuRSaR. However, the cosine distance for FaRoC is slightly lesser as compared to CuRSaR, which indicates that the features extracted using proposed method have higher redundancy compared to that by the CuRSaR.

TABLE I
COMPARATIVE PERFORMANCE ANALYSIS OF DIFFERENT ALGORITHMS FOR BRCA AND OV DATA SETS

Different Data Sets	Different Algorithms	Gene-Protein		Protein-DNA Methylation		Gene-DNA Methylation	
		$B.632+$	Cosine Distance	$B.632+$	Cosine Distance	$B.632+$	Cosine Distance
BRCA	CCA	0.147283	**1.000000** $\pm$ 0.000000	0.157415	**1.000000** $\pm$ 0.000000	0.411243	**1.000000** $\pm$ 0.000000
	RCCA	0.118607	0.999972 $\pm$ 0.000000	0.112662	0.999435 $\pm$ 0.000000	0.132849	0.999991 $\pm$ 0.000000
	SRCCA_{TT}	0.095534	0.999840 $\pm$ 0.000000	0.105675	0.997179 $\pm$ 0.000005	0.066897	0.999961 $\pm$ 0.000000
	SRCCA_{WL}	0.092225	0.999866 $\pm$ 0.000000	0.100061	0.997571 $\pm$ 0.000005	0.064107	0.999743 $\pm$ 0.000000
	SRCCA_{WR}	0.091655	0.999972 $\pm$ 0.000000	0.097832	0.999435 $\pm$ 0.000000	0.064670	0.999974 $\pm$ 0.000000
	SRCCA_{MI}	0.070573	0.999832 $\pm$ 0.000001	0.098611	0.998511 $\pm$ 0.000002	0.068867	0.999977 $\pm$ 0.000000
	SRCCA_{RH}	0.070981	0.999880 $\pm$ 0.000000	0.097450	0.998505 $\pm$ 0.000001	0.071026	0.999961 $\pm$ 0.000000
	CuRSaR	0.071144	0.999850 $\pm$ 0.000000	0.096419	0.996594 $\pm$ 0.000008	0.069176	0.999961 $\pm$ 0.000000
	FaRoC	**0.062137**	0.978141 $\pm$ 0.003873	**0.067812**	0.938846 $\pm$ 0.010906	**0.063843**	0.961144 $\pm$ 0.008728
OV	CCA	0.183022	**1.000000** $\pm$ 0.000000	0.218196	**1.000000** $\pm$ 0.000000	0.211852	**1.000000** $\pm$ 0.000000
	RCCA	0.131289	0.999964 $\pm$ 0.000000	0.143048	**1.000000** $\pm$ 0.000000	0.148742	0.999986 $\pm$ 0.000000
	SRCCA_{TT}	0.129905	0.999236 $\pm$ 0.000000	0.137434	0.999999 $\pm$ 0.000000	0.140221	0.999918 $\pm$ 0.000000
	SRCCA_{WL}	0.129674	0.999747 $\pm$ 0.000000	0.139174	0.996294 $\pm$ 0.000009	0.140199	0.999926 $\pm$ 0.000000
	SRCCA_{WR}	0.129196	0.999473 $\pm$ 0.000000	0.141138	0.994906 $\pm$ 0.000015	0.140232	0.999920 $\pm$ 0.000000
	SRCCA_{MI}	0.129592	0.999543 $\pm$ 0.000000	0.138947	**1.000000** $\pm$ 0.000000	0.139461	0.999733 $\pm$ 0.000000
	SRCCA_{RH}	0.128506	0.999894 $\pm$ 0.000000	0.137111	0.999999 $\pm$ 0.000000	0.139503	0.999863 $\pm$ 0.000000
	CuRSaR	**0.116940**	0.999734 $\pm$ 0.000000	0.124093	0.994793 $\pm$ 0.000017	0.127524	0.999863 $\pm$ 0.000000
	FaRoC	0.119867	0.954086 $\pm$ 0.008265	**0.121938**	0.923905 $\pm$ 0.012295	**0.119630**	0.956203 $\pm$ 0.009183

E. Comparative Performance Analysis

Finally, Fig. 4 and Table I compare the performance of the proposed FaRoC algorithm with that of several existing SRCCA algorithms, namely, SRCCA_{TT} [33], SRCCA_{WR} [33], SRCCA_{WL} [33], SRCCA_{MI}, and SRCCA_{RH}. Results are reported in Fig. 4 for different number of extracted features on three pairs of modalities of two data sets, while Table I compares the minimum $B.632+$ errors obtained using different algorithms. The mean and standard deviation of cosine distance, computed over all extracted features, for each algorithm are also reported in this table for comparison.

From the results reported in Fig. 4 and Table I, it is seen that the performance of the FaRoC algorithm is significantly better than that of other SRCCA algorithms with respect to the $B.632+$ error rate of the SVM. It can also be seen that the SRCCA_{RH} outperforms other SRCCA algorithms in most of the cases. However, the lower values of cosine distance for the proposed algorithm indicate that there exist some redundancy among the features extracted by the FaRoC algorithm.

V. CONCLUSION

This paper presents a new feature extraction algorithm, termed as FaRoC, from two multidimensional data sets. 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 based on spectral decomposition, which helps the proposed FaRoC method to extract required number of correlated features sequentially. The proposed algorithm extracts a new feature from two multidimensional data sets by maximizing its relevance with respect to class label and significance with respect to already-extracted features. The hypercuboid equivalence partition matrix of rough hypercuboid approach is used to compute both the 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 hypercuboid equivalence partition matrix is found to be successful in extracting relevant and significant features from multimodal high dimensional real life data sets. The current formulation shows the utility of rough hypercuboid approach and CCA with respect to knowledge discovery tasks.

REFERENCES

[1] G. R. G. Lanckriet, T. D. Bie, N. Cristianini, M. I. Jordan, and W. S. Noble, "A statistical framework for genomic data fusion," *Bioinformatics*, vol. 20, no. 16, pp. 2626–2635, 2004.

[2] H. Zheng, H. Wang, and D. H. Glass, "Integration of genomic data for inferring protein complexes from global protein–protein interaction networks," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 38, no. 1, pp. 5–16, Feb. 2008.

[3] M. Gönen and E. Alpaydin, "Multiple kernel learning algorithms," *J. Mach. Learn. Res.*, vol. 12, pp. 2211–2268, Jul. 2011.

[4] X. Liu, L. Wang, J. Yin, E. Zhu, and J. Zhang, "An efficient approach to integrating radius information into multiple kernel learning," *IEEE Trans. Cybern.*, vol. 43, no. 2, pp. 557–569, Apr. 2013.

[5] A. Krishnan, L. J. Williams, A. R. McIntosh, and H. Abdi, "Partial least squares (PLS) methods for neuroimaging: A tutorial and review," *NeuroImage*, vol. 56, no. 2, pp. 455–475, 2011.

[6] P. Agius, Y. Ying, and C. Campbell, "Bayesian unsupervised learning with multiple data types," *Stat. Appl. Genet. Molecular Biol.*, vol. 8, no. 1, pp. 1–27, 2009.

[7] I. Huopaniemi, T. Suvitaival, J. Nikkilä, M. Orešič, and S. Kaski, "Multivariate multi-way analysis of multi-source data," *Bioinformatics*, vol. 26, no. 12, pp. i391–i398, 2010.

[8] S. Rogers, A. Klami, J. Sinkkonen, M. Girolami, and S. Kaski, "Infinite factorization of multiple non-parametric views," *Mach. Learn.*, vol. 79, no. 1, pp. 201–226, 2010.

[9] R. S. Savage, Z. Ghahramani, J. E. Griffin, B. J. D. L. Cruz, and D. L. Wild, "Discovering transcriptional modules by Bayesian data integration," *Bioinformatics*, vol. 26, no. 12, pp. i158–i167, 2010.

[10] H. Hotelling, "Relations between two sets of variates," *Biometrika*, vol. 28, nos. 3–4, pp. 321–377, 1936.

[11] W. Zheng, X. Zhou, C. Zou, and L. Zhao, "Facial expression recognition using kernel canonical correlation analysis (KCCA)," *IEEE Trans. Neural Netw.*, vol. 17, no. 1, pp. 233–238, Jan. 2006.

[12] M. Li, Y. Liu, G. Feng, Z. Zhou, and D. Hu, "OI and fMRI signal separation using both temporal and spatial autocorrelations," *IEEE Trans. Biomed. Eng.*, vol. 57, no. 8, pp. 1917–1926, Aug. 2010.

[13] F. Deleus and M. M. V. Hulle, "A connectivity-based method for defining regions-of-interest in fMRI data," *IEEE Trans. Image Process.*, vol. 18, no. 8, pp. 1760–1771, Aug. 2009.

[14] B. Afshin-Pour, S.-M. Shams, and S. Strother, "A hybrid LDA+gCCA model for fMRI data classification and visualization," *IEEE Trans. Med. Imag.*, vol. 34, no. 5, pp. 1031–1041, May 2015.
[15] D. Chu, L.-Z. Liao, M. K. Ng, and X. Zhang, "Sparse canonical correlation analysis: New formulation and algorithm," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 12, pp. 3050–3065, Dec. 2013.
[16] T. Sun and S. Chen, "Locality preserving CCA with applications to data visualization and pose estimation," *Image Vis. Comput.*, vol. 25, no. 5, pp. 531–543, 2007.
[17] K.-A. L. Cao, I. González, and S. Dèjean, "IntegrOmics: An R package to unravel relationships between two Omics datasets," *Bioinformatics*, vol. 25, no. 21, pp. 2855–2856, 2009.
[18] Y. Yamanishi, J. P. Vert, and M. Kanehisa, "Protein network inference from multiple genomic data: A supervised approach," *Bioinformatics*, vol. 20, pp. i363–i370, Aug. 2004.
[19] D. Lin *et al.*, "Group sparse canonical correlation analysis for genomic data integration," *BMC Bioinf.*, vol. 14, p. 245, Aug. 2013.
[20] K. A. L. Cao, P. G. P. Martin, C. R. Granie, and P. Besse, "Sparse canonical methods for biological data integration: Application to a cross-platform study," *BMC Bioinf.*, vol. 10, no. 1, pp. 1–17, 2009.
[21] S. A. Rifkin and J. Kim, "Geometry of gene expression dynamics," *Bioinformatics*, vol. 18, no. 9, pp. 1176–1183, 2002.
[22] L. J. Revell and A. S. Harrison, "PCCA: A program for phylogenetic canonical correlation analysis," *Bioinformatics*, vol. 24, no. 7, pp. 1018–1020, 2008.
[23] A. A. Nielsen, "Multiset canonical correlations analysis and multispectral, truly multitemporal remote sensing data," *IEEE Trans. Image Process.*, vol. 11, no. 3, pp. 293–305, Mar. 2002.
[24] Y.-H. Yuan and Q.-S. Sun, "Multiset canonical correlations using globality preserving projections with applications to feature extraction and recognition," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 25, no. 6, pp. 1131–1146, Jun. 2014.
[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] M. L. Eaton and M. D. Perlman, "The non-singularity of generalized sample covariance matrices," *Ann. Stat.*, vol. 1, no. 4, pp. 710–717, 1973.
[27] I. Gonzalez *et al.*, "Highlighting relationships between heterogeneous biological data through graphical displays based on regularized canonical correlation analysis," *J. Biol. Syst.*, vol. 17, no. 2, pp. 173–199, 2009.
[28] H. D. Vinod, "Canonical ridge and econometrics of joint production," *J. Econometrics*, vol. 4, no. 2, pp. 147–166, 1976.
[29] I. González, S. Dèjean, P. G. P. Martin, and A. Baccini, "CCA: An R package to extend canonical correlation analysis," *J. Stat. Softw.*, vol. 23, no. 12, pp. 1–14, 2008.
[30] T. D. Bie and B. D. Moor, "On the regularization of canonical correlation analysis," in *Proc. 4th Int. Symp. Independent Compon. Anal. Blind Signal Separation*, Nara, Japan, 2003, pp. 785–790.
[31] A. E. Hoerl and R. W. Kennard, "Ridge regression: Biased estimation for nonorthogonal problems," *Technometrics*, vol. 12, no. 1, pp. 55–67, 1970.
[32] M. Kang *et al.*, "eQTL mapping study via regularized sparse canonical correlation analysis," in *Proc. 12th Int. Conf. Mach. Learn. Appl.*, Miami, FL, USA, 2013, pp. 129–134.
[33] A. Golugula *et al.*, "Supervised regularized canonical correlation analysis: Integrating histologic and proteomic measurements for predicting biochemical recurrence following prostate surgery," *BMC Bioinf.*, vol. 12, pp. 483–495, Dec. 2011.
[34] P. Maji and A. Mandal, "Multimodal omics data integration using max relevance-max significance criterion," *IEEE Trans. Biomed. Eng.*, to be published, doi: 10.1109/TBME.2016.2624823.
[35] Z. Pawlak, *Rough Sets: Theoretical Aspects of Reasoning About Data*. Dordrecht, The Netherlands: Kluwer, 1991.
[36] P. Maji and S. K. Pal, *Rough-Fuzzy Pattern Recognition: Applications in Bioinformatics and Medical Imaging*. Hoboken, NJ, USA: Wiley, 2012.
[37] P. Maji, "Fuzzy–rough supervised attribute clustering algorithm and classification of microarray data," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 41, no. 1, pp. 222–233, Feb. 2011.
[38] P. Maji and S. Paul, "Rough set based maximum relevance-maximum significance criterion and gene selection from microarray data," *Int. J. Approx. Reason.*, vol. 52, no. 3, pp. 408–426, 2011.
[39] P. Maji and S. K. Pal, "Fuzzy–rough sets for information measures and selection of relevant genes from microarray data," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 40, no. 3, pp. 741–752, Jun. 2010.
[40] S. Paul and P. Maji, "Rough sets for in silico identification of differentially expressed miRNAs," *Int. J. Nanomed.*, vol. 8, pp. 63–74, Sep. 2013.
[41] P. Maji and S. Paul, "Robust rough-fuzzy C-means algorithm: Design and applications in coding and non-coding RNA expression data clustering," *Fundamenta Informaticae*, vol. 124, nos. 1–2, pp. 153–174, 2013.
[42] S. Paul and P. Maji, "City block distance and rough-fuzzy clustering for identification of co-expressed microRNAs," *Mol. BioSyst.*, vol. 10, no. 6, pp. 1509–1523, 2014.
[43] P. Maji and S. Paul, "Rough-fuzzy clustering for grouping functionally similar genes from microarray data," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 10, no. 2, pp. 286–299, Mar./Apr. 2013.
[44] P. Maji, "A rough hypercuboid approach for feature selection in approximation spaces," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 1, pp. 16–29, Jan. 2014.
[45] S. Paul and P. Maji, "μHEM for identification of differentially expressed miRNAs using hypercuboid equivalence partition matrix," *BMC Bioinf.*, vol. 14, no. 1, p. 266, 2013.
[46] P. Maji and A. Mandal, "Rough hypercuboid based supervised regularized canonical correlation for multimodal data analysis," *Fundamenta Informaticae*, vol. 148, nos. 1–2, pp. 133–155, 2016.
[47] G. M. L. Gladwell, "On isospectral spring-mass systems," *Inverse Problems*, vol. 11, no. 3, pp. 591–602, 1995.
[48] Y. Guo, T. Hastie, and R. Tibshirani, "Regularized linear discriminant analysis and its application in microarrays," *Biostatistics*, vol. 8, no. 1, pp. 86–100, 2007.
[49] N. H. Timm, *Applied Multivariate Analysis*. New York, NY, USA: Springer, 2002.
[50] E. J. Ientilucci, "Using the singular value decomposition," Chester F. Carlson Center Imag. Sci., Rochester Inst. Technol., Rochester, NY, USA, 2003.
[51] K. S. Miller, "On the Inverse of the sum of matrices," *Math. Mag.*, vol. 54, no. 2, pp. 67–72, 1981.
[52] G. H. Golub and C. F. V. Loan, *Matrix Computations*, 3rd ed. Baltimore, MD, USA: Johns Hopkins Univ. Press, 1996.
[53] P. A. White, "The computation of eigenvalues and eigenvectors of a matrix," *J. Soc. Ind. Appl. Math.*, vol. 6, no. 4, pp. 393–437, 1958.
[54] B. Efron and R. Tibshirani, "Improvements on cross-validation: The .632+ bootstrap method," *J. Amer. Stat. Assoc.*, vol. 92, no. 438, pp. 548–560, 1997.
[55] V. N. Vapnik, *The Nature of Statistical Learning Theory*. New York, NY, USA: Springer, 1995.

Ankita Mandal received the B.Sc. degree in electronics from West Bengal State University, Kolkata, India, in 2011, and the Master of Computer Application degree from Jadavpur University, Kolkata, in 2014.

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

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 with the Machine Intelligence Unit, Indian Statistical Institute, Kolkata. He has published over 120 papers in international journals and conferences. He has authored a book published by Wiley-IEEE Computer Society Press and another book published by Springer-Verlag, London. His current research interests include pattern recognition, machine learning, computational biology and bioinformatics, and medical image processing.

Dr. Maji was a recipient of 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.