Rough Hypercuboid Approach for Feature Selection in Approximation Spaces

Pradipta Maji

Abstract—The selection of relevant and significant features is an important problem particularly for data sets with large number of features. In this regard, a new feature selection algorithm is presented based on rough hypercuboid approach. It selects a set of features from a data set by maximizing the relevance, dependency, and significance of the selected features. By introducing the concept of hypercuboid equivalence partition matrix, a novel representation of degree of dependency of sample categories on features is proposed to measure the relevance, dependency, and significance of features in approximation spaces. The equivalence partition matrix also offers an efficient way to calculate many more quantitative measures to describe the inexactness of approximate classification. Several quantitative indices are introduced based on rough hypercuboid approach for evaluating the performance of proposed method. The superiority of the proposed method over other feature selection methods, in terms of computational complexity and classification accuracy, is established extensively on various real life data sets of different sizes and dimensions.

Index Terms—Pattern recognition, data mining, feature selection, rough sets, rough hypercuboid approach.

1 Introduction

Dimensionality reduction or feature selection from a data set is an essential preprocessing step used for mining large data sets, both in dimension and size [1], [2]. Many problems in pattern recognition, data mining, and machine learning may involve thousands of features. One of the important problems of large data analysis is to obtain a smaller set of representative features by preserving the semantics of the data. It leads to more compactness of the models learned and better generalization as well as decreases the processing time. Hence, the objective is to reduce dimensionality using information contained within the data set and preserve most relevant information of the original data according to some optimality criteria [1], [2].

The optimal characterization condition often means the minimal classification error, which usually requires the maximal statistical dependency of the sample categories or class labels on the data distribution in the reduced feature space. This scheme is called maximal dependency or Max-Dependency, in which, the task of feature selection is to find a feature subset from the whole feature set, which jointly have the largest dependency on the target class.

An optimal feature subset selected by a dimensionality reduction method is always relative to a certain feature evaluation criterion. In general, different criteria may lead to different optimal feature subsets. However, every criterion tries to measure the discriminating ability of a feature or a subset of features to distinguish different class labels. One of the main problems in real life data analysis is uncertainty. Some of the sources of this uncertainty include incompleteness and vagueness in class definitions. In this background, the possibility concept introduced by rough set theory has gained popularity in modeling and propagating uncertainty [3], [4]. It has been successfully used to find informative feature subset from the original attributes of a data set with discretized attribute values [5], [6], [7], [8], [9], [10]. The quick reduct algorithm of Chouchoulas and Shen [7] is based on the principle of Max-Dependency criterion.

There are usually real valued data and fuzzy information in real world applications. In rough set theory, the real valued features are divided into several discrete partitions to calculate the dependency of a feature. However, a best partition of feature values is NP-hard problem [11]. Also, the inherent error that exists in discretization process is of major concern in the computation of the dependency of real valued features. Combining fuzzy and rough sets provides an important direction in reasoning with uncertainty for real valued data sets [12], [13], [14], [15], [16]. The generalized theory of rough-fuzzy computing has been applied successfully to feature selection of real valued data [16], [17], [18], [19], [20], [21], [22]. Also, neighborhood rough sets [23] and rough hypercuboid approach [24] are found to be suitable for numerical data sets. The fuzzy-rough quick reduct algorithm due to Jensen and Shen [17] and neighborhood rough set based feature selection algorithm of Hu et al. [23] are based on Max-Dependency criterion.

However, for real life high dimensional data set, the number of samples is often inadequate and generation of equivalence classes for rough sets is usually an ill-posed problem. In effect, it is very hard to estimate correctly the joint dependency of the features when the number of equivalence classes increases very quickly and gets comparable to the number of samples. Also, the computational speed of Max-Dependency is very

• *The author is with the Machine Intelligence Unit, Indian Statistical Institute, Kolkata, India. E-mail: pmaji@isical.ac.in.*

slow. Hence, Max-Dependency feature selection is not appropriate for real life applications where the aim is to achieve high classification accuracy with a reasonably compact set of features, although it might be useful to select a very small number of features [25].

An alternative to Max-Dependency criterion is to select features based on maximal relevance or Max-Relevance criterion. It is to search a set of features that approximates Max-Dependency criterion with the mean value of all dependency values between individual feature and target class label [26], [27], [28]. However, unlike Max-Dependency criterion, this criterion does not consider the joint effect of features on the target class. Also, it is likely that features selected according to Max-Relevance could have rich redundancy. The minimal redundancy (Min-Redundancy) or maximal significance (Max-Significance) criterion can be used for searching mutually exclusive or independent features. However, this is not sufficient for selecting highly discriminative features. Combining redundancy or significance criterion with relevance criterion, the minimal redundancy-maximal relevance (mRMR) [29] and maximal relevance-maximal significance (MRMS) [25] criteria have been proposed to select relevant and nonredundant or significant features, although both mRMR and MRMS criteria do not consider dependency between the data distribution in multidimensional space and class labels.

In this paper, a novel feature selection method is proposed, which employs rough hypercuboid approach to provide a means by which real valued noisy data can be effectively reduced without the need for user-specified information. The proposed method selects a subset of features from whole feature set by maximizing relevance, dependency, and significance of the selected features. Using the novel concept of hypercuboid equivalence partition matrix, the degree of dependency is calculated for condition attributes, which is used to compute relevance, dependency, and significance of the features. Hence, the only information required in the proposed method is in the form of equivalence classes for each attribute, which can be automatically derived from the data set. Several quantitative measures are introduced based on rough hypercuboid approach to evaluate the performance of proposed feature selection method. The effectiveness of the proposed method, along with a comparison with other methods, is demonstrated on a set of real life data.

The structure of this paper is as follows: Section 2 briefly introduces the necessary notions of rough sets. In Section 3, the formulae of degree of dependency is introduced for approximation spaces with a hypercuboid equivalence partition matrix. The proposed feature selection method based on hypercuboid equivalence partition matrix is described in Section 4. Different quantitative measures are introduced in Section 5 based on the concept of rough hypercuboid to evaluate the inexactness of approximate classification. A few case studies and a comparison with other methods are presented in Section 6. Concluding remarks are given in Section 7.

2 Rough Sets

An approximation space or information system is a pair $< \mathbb{U}, \mathbb{A} >$ [3], where $\mathbb{U} = \{x_1, \cdots, x_i, \cdots, x_n\}$ be a non-empty set, the universe of discourse, and $\mathbb{A}$ is a family of attributes, also called knowledge in the universe. V is the value domain of $\mathbb{A}$ and f is an information function $f : \mathbb{U} \times \mathbb{A} \rightarrow V$. Any subset $\mathbb{P}$ of knowledge $\mathbb{A}$ defines an equivalence or indiscernibility relation $IND(\mathbb{P})$ on $\mathbb{U}$

$$IND(\mathbb{P}) = \{(x_i, x_j) \in \mathbb{U} \times \mathbb{U} | \forall a \in \mathbb{P}, f(x_i, a) = f(x_j, a)\}.$$

If $(x_i, x_j) \in IND(\mathbb{P})$, then x_i and x_j are indiscernible by attributes from $\mathbb{P}$. The partition of $\mathbb{U}$ generated by $IND(\mathbb{P})$ is denoted as

$$\mathbb{U}/IND(\mathbb{P}) = \{[x_i]_{\mathbb{P}} : x_i \in \mathbb{U}\} \quad (1)$$

where $[x_i]_{\mathbb{P}}$ is the equivalence class containing x_i. The elements in $[x_i]_{\mathbb{P}}$ are indiscernible or equivalent with respect to knowledge $\mathbb{P}$. Equivalence classes, also termed as information granules, are used to characterize arbitrary subsets of $\mathbb{U}$. The equivalence classes of $IND(\mathbb{P})$ and the empty set $\emptyset$ are the elementary sets in the approximation space $< \mathbb{U}, \mathbb{A} >$.

Given an arbitrary set $X \subseteq \mathbb{U}$, in general, it may not be possible to describe X precisely in $< \mathbb{U}, \mathbb{A} >$. One may characterize X by a pair of lower and upper approximations, defined as follows [3]:

$$\underline{\mathbb{P}}(X) = \bigcup \{[x_i]_{\mathbb{P}} \mid [x_i]_{\mathbb{P}} \subseteq X\} \text{ and} \quad (2)$$

$$\overline{\mathbb{P}}(X) = \bigcup \{[x_i]_{\mathbb{P}} \mid [x_i]_{\mathbb{P}} \cap X \neq \emptyset\}. \quad (3)$$

Hence, the lower approximation $\underline{\mathbb{P}}(X)$ is the union of all the elementary sets which are subsets of X, and the upper approximation $\overline{\mathbb{P}}(X)$ is the union of all the elementary sets which have a non-empty intersection with X. The tuple $< \underline{\mathbb{P}}(X), \overline{\mathbb{P}}(X) >$ is the representation of an ordinary set X in the approximation space $< \mathbb{U}, \mathbb{A} >$ or simply called the rough set of X. The lower (respectively, upper) approximation $\underline{\mathbb{P}}(X)$ (respectively, $\overline{\mathbb{P}}(X)$) is interpreted as the collection of those elements of $\mathbb{U}$ that definitely (respectively, possibly) belong to X. The lower approximation is also called positive region sometimes, denoted as $POS_{\mathbb{P}}(X)$. A set X is said to be definable or exact in $< \mathbb{U}, \mathbb{A} >$ iff $\underline{\mathbb{P}}(X) = \overline{\mathbb{P}}(X)$. Otherwise X is indefinable and termed as a rough set. $BND_{\mathbb{P}}(X) = \overline{\mathbb{P}}(X) \setminus \underline{\mathbb{P}}(X)$ is called a boundary set.

Definition 1: An information system $< \mathbb{U}, \mathbb{A} >$ is called a decision table if the set $\mathbb{A} = \mathbb{C} \cup \mathbb{D}$, where $\mathbb{C}$ and $\mathbb{D}$ are condition and decision attribute sets, respectively. The dependency between $\mathbb{C}$ and $\mathbb{D}$ can be defined as [3]

$$\gamma_{\mathbb{C}}(\mathbb{D}) = \frac{|POS_{\mathbb{C}}(\mathbb{D})|}{|\mathbb{U}|} \quad (4)$$

where $POS_{\mathbb{C}}(\mathbb{D}) = \cup \underline{\mathbb{C}} X_i$, X_i is the ith equivalence class induced by $\mathbb{D}$ and $|\cdot|$ denotes the cardinality of a set.

Definition 2: Given $\mathbb{C}, \mathbb{D}$ and an attribute $\mathcal{A} \in \mathbb{C}$, the significance of the attribute $\mathcal{A}$ is defined as [3]:

$$\sigma_{\mathbb{C}}(\mathbb{D}, \mathcal{A}) = \gamma_{\mathbb{C}}(\mathbb{D}) - \gamma_{\mathbb{C}-\{\mathcal{A}\}}(\mathbb{D}). \quad (5)$$

3 Hypercuboid Partition Matrix

In this section, the concept of hypercuboid equivalence partition matrix, based on rough hypercuboid approach [24], is introduced to compute the degree of dependency of decision attribute set on the condition attribute set.

3.1 Hypercuboid

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.

Fig. 1 presents the scatter plots of samples from two classes, namely, Class A and Class B, considering two attributes $\mathcal{A}_i$ and $\mathcal{A}_j$. The intervals $[\mathrm{A}, \mathrm{C}]$ and $[\mathrm{B}, \mathrm{D}]$ are the value ranges of attribute $\mathcal{A}_i$ with respect to Class A and Class B, respectively. That is, the attribute $\mathcal{A}_i$ value of each sample with class label A falls within the interval $[\mathrm{A}, \mathrm{C}]$, while that with class label B belongs within the interval $[\mathrm{B}, \mathrm{D}]$. These two intervals form two hypercuboids along the attribute $\mathcal{A}_i$. Similarly, two hypercuboids can be formed by the intervals $[\mathrm{E}, \mathrm{G}]$ and $[\mathrm{F}, \mathrm{H}]$ corresponding to Class A and Class B, respectively, along the attribute $\mathcal{A}_j$.

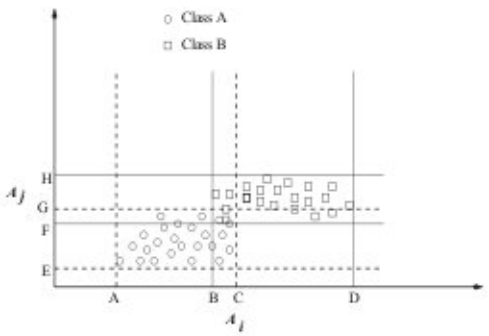

Fig. 1. Rough hypercuboids in two dimension

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. Hence, the 2-dimensional hypercuboid for Class A can be formed by taking the Cartesian product $[\mathrm{A}, \mathrm{C}] \times [\mathrm{E}, \mathrm{G}]$ of two orthogonal intervals $[\mathrm{A}, \mathrm{C}]$ and $[\mathrm{E}, \mathrm{G}]$ corresponding to the attributes $\mathcal{A}_i$ and $\mathcal{A}_j$, respectively. This hypercuboid is also referred to as the class hypercuboid of Class A. Similarly, the Cartesian product $[\mathrm{B}, \mathrm{D}] \times [\mathrm{F}, \mathrm{H}]$ of two orthogonal intervals $[\mathrm{B}, \mathrm{D}]$ and $[\mathrm{F}, \mathrm{H}]$ forms the 2-dimensional hypercuboid for Class B. For all hypercuboids, any two objects belong to a same class hypercuboid are said to be indiscernible with respect to that particular class.

However, in real data analysis, uncertainty arises due to overlapping class boundaries. Hence, every two class hypercuboids may intersect with each other. The intersection of two hypercuboids also forms a hypercuboid, which is referred to as implicit hypercuboid. The implicit hypercuboids encompass the misclassified samples or objects those belong to more than one classes. The degree of dependency of the decision attribute set or class label on the condition attribute set depends on the cardinality of the implicit hypercuboids. The degree of dependency increases with the decrease in cardinality.

3.2 Hypercuboid Equivalence Partition Matrix

Let $\mathbb{U} = \{x_1, \cdots, x_j, \cdots, x_n\}$ be the finite set of n objects, and $\mathbb{C} = \{\mathcal{A}_1, \cdots, \mathcal{A}_k, \cdots, \mathcal{A}_m\}$ and $\mathbb{D}$ are the condition and decision attribute sets in $\mathbb{U}$, respectively. If $\mathbb{U}/\mathbb{D} = \{\beta_1, \cdots, \beta_i, \cdots, \beta_c\}$ denotes c equivalence classes or information granules of $\mathbb{U}$ generated by the equivalence relation induced from the decision attribute set $\mathbb{D}$, then c equivalence classes of $\mathbb{U}$ can also be generated by the equivalence relation induced from each condition attribute $\mathcal{A}_k \in \mathbb{C}$. If $\mathbb{U}/\mathcal{A}_k = \{\delta_1, \cdots, \delta_i, \cdots, \delta_c\}$ denotes c equivalence classes or information granules of $\mathbb{U}$ induced by the condition attribute $\mathcal{A}_k$ and n is the number of objects in $\mathbb{U}$, then c-partitions of $\mathbb{U}$ are the sets of (cn) values $\{\mathrm{h}_{ij}(\mathcal{A}_k)\}$ that can be conveniently arrayed as a $(c \times n)$ matrix $\mathbb{H}(\mathcal{A}_k) = [\mathrm{h}_{ij}(\mathcal{A}_k)]$. The matrix $\mathbb{H}(\mathcal{A}_k)$ is termed as hypercuboid equivalence partition matrix of the condition attribute $\mathcal{A}_k$ and is denoted by

$$\mathbb{H}(\mathcal{A}_k) = \begin{pmatrix} \mathrm{h}_{11}(\mathcal{A}_k) & \mathrm{h}_{12}(\mathcal{A}_k) & \cdots & \mathrm{h}_{1n}(\mathcal{A}_k) \\ \mathrm{h}_{21}(\mathcal{A}_k) & \mathrm{h}_{22}(\mathcal{A}_k) & \cdots & \mathrm{h}_{2n}(\mathcal{A}_k) \\ \cdots & \cdots & \cdots & \cdots \\ \mathrm{h}_{c1}(\mathcal{A}_k) & \mathrm{h}_{c2}(\mathcal{A}_k) & \cdots & \mathrm{h}_{cn}(\mathcal{A}_k) \end{pmatrix} \tag{6}$$

$$\text{where } \mathrm{h}_{ij}(\mathcal{A}_k) = \begin{cases} 1 & \text{if } \mathrm{L}_i \le x_j(\mathcal{A}_k) \le \mathrm{U}_i \\ 0 & \text{otherwise.} \end{cases} \tag{7}$$

Here $\mathrm{h}_{ij}(\mathcal{A}_k) \in \{0, 1\}$ represents the membership of object x_j in the ith equivalence partition or class β_i satisfying following two conditions:

$$1 \le \sum_{j=1}^{n} \mathrm{h}_{ij}(\mathcal{A}_k) \le n, \forall i; \quad 1 \le \sum_{i=1}^{c} \mathrm{h}_{ij}(\mathcal{A}_k) \le c, \forall j. \tag{8}$$

The above axioms should hold for every equivalence partition, which correspond to the requirement that an equivalence class is non-empty. The tuple $[\mathrm{L}_i, \mathrm{U}_i]$ represents the interval of ith class β_i according to the decision attribute set $\mathbb{D}$. The interval $[\mathrm{L}_i, \mathrm{U}_i]$ 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 x_j with class label β_i falls within interval $[\mathrm{L}_i, \mathrm{U}_i]$. This can be viewed as a supervised granulation process, which utilizes class information.

In general, the number of class hypercuboids is equal to the number of classes c. Each hypercuboid corresponds to a unique class. However, more than one class hypercuboids may correspond to one class. If the interval

$[\mathrm{L}_i, \mathrm{U}_i]$ corresponding to class β_i is completely covered by another interval $[\mathrm{L}_j, \mathrm{U}_j]$ of class β_j and no object of class β_j falls within the interval $[\mathrm{L}_i, \mathrm{U}_i]$, the interval $[\mathrm{L}_j, \mathrm{U}_j]$ is then cut and shrunken into two short intervals for class β_j, namely, $[\mathrm{L}_j, \mathrm{L}_i]$ and $[\mathrm{U}_i, \mathrm{U}_j]$, both of which don't overlap with the interval $[\mathrm{L}_i, \mathrm{U}_i]$ of class β_i.

However, such a granulation process does not necessarily result in a compatible granulation in the sense that every two intervals may intersect with each other. These intersections form the implicit hypercuboids. The degree of dependency of decision attribute on a condition attribute or a subset of attributes is evaluated by finding the implicit hypercuboids that encompass misclassified objects. Using the concept of hypercuboid equivalence partition matrix, the misclassified objects of implicit hypercuboids can be identified based on the confusion vector defined next

$$\mathbb{V}(\mathcal{A}_k) = [\mathrm{v}_1(\mathcal{A}_k), \cdots, \mathrm{v}_j(\mathcal{A}_k), \cdots, \mathrm{v}_n(\mathcal{A}_k)] \tag{9}$$

$$\text{where} \quad \mathrm{v}_j(\mathcal{A}_k) = \min\{1, \sum_{i=1}^{c} \mathrm{h}_{ij}(\mathcal{A}_k) - 1\}. \tag{10}$$

In other words, if an object x_j belongs to the lower approximation of any class β_i, then it does not belong to the lower or upper approximations of any other classes and $\mathrm{v}_j(\mathcal{A}_k) = 0$. On the other hand, if the object x_j belongs to the boundary region of more than one classes, then it should be encompassed by the implicit hypercuboid and $\mathrm{v}_j(\mathcal{A}_k) = 1$. Hence, the hypercuboid equivalence partition matrix and corresponding confusion vector of the condition attribute $\mathcal{A}_k$ can be used to define the lower and upper approximations of the ith class β_i of the decision attribute set $\mathbb{D}$.

Let $\beta_i \subseteq \mathbb{U}$. β_i can be approximated using only the information contained within $\mathcal{A}_k$ by constructing the A-lower and A-upper approximations of β_i:

$$\underline{A}(\beta_i) = \{x_j |\ \mathrm{h}_{ij}(\mathcal{A}_k) = 1 \text{ and } \mathrm{v}_j(\mathcal{A}_k) = 0\}; \tag{11}$$

$$\overline{A}(\beta_i) = \{x_j |\ \mathrm{h}_{ij}(\mathcal{A}_k) = 1\}; \tag{12}$$

where equivalence relation A is induced from attribute $\mathcal{A}_k$. The boundary region of β_i is then defined as

$$BND_A(\beta_i) = \overline{A}(\beta_i) \setminus \underline{A}(\beta_i) \tag{13}$$

$$\text{i.e., } BND_A(\beta_i) = \{x_j |\ \mathrm{h}_{ij}(\mathcal{A}_k) = 1 \text{ and } \mathrm{v}_j(\mathcal{A}_k) = 1\}. \tag{14}$$

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 partition or class. The ith hypercuboid equivalence partition is, therefore, given by

$$\beta_i = \{\mathrm{h}_{i1}(\mathcal{A}_k)/x_1 + \mathrm{h}_{i2}(\mathcal{A}_k)/x_2 + \cdots + \mathrm{h}_{in}(\mathcal{A}_k)/x_n\}. \tag{15}$$

As to a hypercuboid partition induced by an equivalence relation, the equivalence class is a set. "+" means the operator of union in this case. The cardinality of the set is the cardinality of the upper approximation of the class β_i that can be calculated with

$$|\overline{A}(\beta_i)| = \sum_{j=1}^{n} \mathrm{h}_{ij}(\mathcal{A}_k) \tag{16}$$

which appears to be a natural generalization of crisp set. Similarly, the cardinalities of lower approximation and boundary region of class β_i can be calculated as follows:

$$|\underline{A}(\beta_i)| = \sum_{j=1}^{n} \mathrm{h}_{ij}(\mathcal{A}_k) \cap [1 - \mathrm{v}_j(\mathcal{A}_k)]; \tag{17}$$

$$|BND_A(\beta_i)| = \sum_{j=1}^{n} \mathrm{h}_{ij}(\mathcal{A}_k) \cap \mathrm{v}_j(\mathcal{A}_k). \tag{18}$$

3.3 Dependency and Significance

Based on the definitions of lower and upper approximations, the positive, negative, and boundary regions of decision attribute set $\mathbb{D}$ can be defined as:

$$POS_A(\mathbb{D}) = \bigcup_{\beta_i \in \mathbb{U}/\mathbb{D}} \underline{A}(\beta_i); \tag{19}$$

$$NEG_A(\mathbb{D}) = \mathbb{U} - \bigcup_{\beta_i \in \mathbb{U}/\mathbb{D}} \overline{A}(\beta_i); \tag{20}$$

$$BND_A(\mathbb{D}) = \bigcup_{\beta_i \in \mathbb{U}/\mathbb{D}} \overline{A}(\beta_i) - \bigcup_{\beta_i \in \mathbb{U}/\mathbb{D}} \underline{A}(\beta_i). \tag{21}$$

The positive region, $POS_A(\mathbb{D})$, contains all objects of $\mathbb{U}$ that can be classified to classes of $\mathbb{U}/\mathbb{D}$ using the knowledge in attribute $\mathcal{A}_k$. The boundary region, $BND_A(\mathbb{D})$, is the set of objects that can possibly, but not certainly, be classified in this way. The negative region, $NEG_A(\mathbb{D})$, is the set of objects that cannot be classified to classes of $\mathbb{U}/\mathbb{D}$.

Combining (6), (9), and (19), 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

$$|POS_A(\mathbb{D})| = \sum_{i=1}^{c} \sum_{j=1}^{n} \mathrm{h}_{ij}(\mathcal{A}_k) \cap [1 - \mathrm{v}_j(\mathcal{A}_k)]. \tag{22}$$

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

$$\gamma_{\mathcal{A}_k}(\mathbb{D}) = \frac{1}{n} \sum_{i=1}^{c} \sum_{j=1}^{n} \mathrm{h}_{ij}(\mathcal{A}_k) \cap [1 - \mathrm{v}_j(\mathcal{A}_k)], \tag{23}$$

$$\text{that is, } \gamma_{\mathcal{A}_k}(\mathbb{D}) = 1 - \frac{1}{n} \sum_{j=1}^{n} \mathrm{v}_j(\mathcal{A}_k), \tag{24}$$

where $0 \leq \gamma_{\mathcal{A}_k}(\mathbb{D}) \leq 1$. 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 $\mathcal{A}_k$.

An important issue in real life data analysis is discovering dependencies between attributes. Intuitively, a set of attributes $\mathbb{D}$ depends totally on a set of attributes $\mathbb{C}$, denoted $\mathbb{C} \Rightarrow \mathbb{D}$, if all attribute values from $\mathbb{D}$ are

4.2 Computational Complexity

Prior to computing the relevance, dependency, or significance of a condition attribute, the hypercuboid equivalence partition matrix and confusion vector for each condition attribute are to be generated first.

The computational complexity to generate a $(c \times n)$ hypercuboid equivalence partition matrix is $\mathcal{O}(cn)$, where c and n represent the number of classes and objects in the data set, respectively, while the generation of confusion vector has $\mathcal{O}(n)$ time complexity. In effect, the computation of the relevance of a feature has $(\mathcal{O}(cn+n) =)\mathcal{O}(cn)$ time complexity. Hence, the total complexity to compute the relevance of m features, which is carried out in step 2 of the proposed algorithm, is $\mathcal{O}(mcn)$. The selection of most relevant feature from the set of m features, which is carried out in step 3, has a complexity $\mathcal{O}(m)$.

There is only one loop in step 4 of the proposed feature selection method, which is executed $(d-1)$ times, where d represents the number of selected features. The computation of dependency of a candidate feature has the complexity $\mathcal{O}(cn)$. Similarly, the complexity to compute the significance of a candidate feature with respect to another feature has also the complexity $\mathcal{O}(cn)$. If $\acute{m}$ represents the cardinality of the already selected feature set, the total complexity to compute the significance and dependency of $(m - \acute{m})$ candidate features, which is carried out in step 5, is $\mathcal{O}((m-\acute{m})cn)$. The selection of a feature from $(m-\acute{m})$ candidate features by maximizing relevance, dependency, and significance, which is carried out in step 6, has a complexity $\mathcal{O}(m - \acute{m})$. Hence, the total complexity to execute the loop $(d - 1)$ times is $(\mathcal{O}((d-1)((m-\acute{m}) + (m - \acute{m})cn)) =)\mathcal{O}(dcn(m - \acute{m}))$.

In effect, the selection of a set of d relevant and significant features from the whole set of m features using the proposed hypercuboid equivalence partition matrix based first order incremental search method has an overall computational complexity of $(\mathcal{O}(mcn) + \mathcal{O}(m) + \mathcal{O}(dcn(m - \acute{m})) =)\mathcal{O}(dnm)$ as $c, \acute{m} << m$.

5 Quantitative Indices

In this section, some quantitative indices are introduced, incorporating the concepts of rough sets [3] and hypercuboid equivalence partition matrix.

5.1 Average Accuracy, α Index

The α index represents the average accuracy of c classes. It is the average of the ratio of number of objects in lower approximation to that in upper approximation of each class and is given by

$$\alpha = \frac{1}{c}\sum_{i=1}^{c}\frac{|\underline{A}(\beta_i)|}{|\overline{A}(\beta_i)|} \tag{36}$$

$$\text{where } |\overline{A}(\beta_i)| = \sum_{j=1}^{n} \mathrm{h}_{ij}(\mathbb{S}); \tag{37}$$

$$|\underline{A}(\beta_i)| = \sum_{j=1}^{n} \mathrm{h}_{ij}(\mathbb{S}) \cap [1 - \mathrm{v}_j(\mathbb{S})] \tag{38}$$

represent the cardinalities of upper and lower approximations of the class β_i, respectively, $\mathrm{h}_{ij}(\mathbb{S})$ and $\mathrm{v}_j(\mathbb{S})$ are hypercuboid equivalence partition matrix and confusion vector of condition attribute set $\mathbb{S}$, respectively.

The α index captures the average degree of completeness of knowledge about all classes. A good feature set should make all classes as separable as possible. The α index increases with decrease in overlapping among different classes. When $\overline{A}(\beta_i) = \underline{A}(\beta_i)$, $\forall i$, that is, all the classes $\{\beta_i\}$ are exact or definable, then we have $\alpha = 1$. Whereas if $\overline{A}(\beta_i) = BND_A(\beta_i)$, $\forall i$, the value of $\alpha = 0$. Hence, $0 \leq \alpha \leq 1$.

5.2 Average Roughness, ϱ Index

It represents the average roughness of c classes and is defined by subtracting the average accuracy α from 1:

$$\varrho = 1 - \alpha = 1 - \frac{1}{c}\sum_{i=1}^{c}\frac{|\underline{A}(\beta_i)|}{|\overline{A}(\beta_i)|} \tag{39}$$

where $|\overline{A}(\beta_i)|$ and $|\underline{A}(\beta_i)|$ are given by (37) and (38), respectively. Note that the lower the value of ϱ, the better is the overall classes approximations. Also, $0 \leq \varrho \leq 1$. Basically, ϱ index represents the average degree of incompleteness of knowledge about all classes.

5.3 Accuracy of Approximation, $\alpha^\star$ Index

The $\alpha^\star$ index represents the accuracy of approximation of all classes and can be defined as

$$\alpha^\star = \frac{\sum_{i=1}^{c}|\underline{A}(\beta_i)|}{\sum_{i=1}^{c}|\overline{A}(\beta_i)|} \tag{40}$$

where $|\overline{A}(\beta_i)|$ and $|\underline{A}(\beta_i)|$ are given by (37) and (38), respectively. It captures the exactness of approximate classification. A good feature selection procedure should make the value of $\alpha^\star$ as high as possible. The $\alpha^\star$ index maximizes the exactness of approximate classification.

5.4 Quality of Approximation, γ Index

It is the ratio of the total number of objects in lower approximations of all classes to the cardinality of the universe of discourse $\mathbb{U}$ and is given by

$$\gamma = \frac{1}{|\mathbb{U}|}\sum_{i=1}^{c}|\underline{A}(\beta_i)|, \tag{41}$$

where $|\underline{A}(\beta_i)|$ is given by (38). The γ index basically represents the quality of approximation of a classification algorithm. A good feature selection procedure should make the value of γ index as high as possible.

uniquely determined by values of attributes from $\mathbb{C}$. If there exists a functional dependency between values of $\mathbb{C}$ and $\mathbb{D}$, then $\mathbb{D}$ depends totally on $\mathbb{C}$.

Given $< \mathbb{U}, \mathbb{A} >$, $\mathcal{A}_k$ and $\mathcal{A}_l$ are two condition attributes of $\mathbb{A}$. The $c \times n$ hypercuboid equivalence partition matrix corresponding to the set $\{\mathcal{A}_k, \mathcal{A}_l\}$ 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:

$$\mathbb{H}(\{\mathcal{A}_k, \mathcal{A}_l\}) = \mathbb{H}(\mathcal{A}_k) \cap \mathbb{H}(\mathcal{A}_l); \tag{25}$$

$$\text{where } \mathrm{h}_{ij}(\{\mathcal{A}_k, \mathcal{A}_l\}) = \mathrm{h}_{ij}(\mathcal{A}_k) \cap \mathrm{h}_{ij}(\mathcal{A}_l). \tag{26}$$

Hence, the $c \times n$ hypercuboid equivalence partition matrix of the set $\mathbb{C} = \{\mathcal{A}_1, \cdots, \mathcal{A}_k, \cdots, \mathcal{A}_m\}$ of condition attributes is given by

$$\mathbb{H}(\mathbb{C}) = \bigcap_{\mathcal{A}_k \in \mathbb{C}} \mathbb{H}(\mathcal{A}_k); \quad \text{where } \mathrm{h}_{ij}(\mathbb{C}) = \bigcap_{\mathcal{A}_k \in \mathbb{C}} \mathrm{h}_{ij}(\mathcal{A}_k). \tag{27}$$

In this regard, it should be mentioned that the concept of positive approximation accelerator [30] can be used to improve the computing performance of dependency of a condition attribute set $\mathbb{C}_{i+1} = \{\mathcal{A}_1, \mathcal{A}_2, \cdots, \mathcal{A}_{i+1}\}$ using the following recursive expression principle:

$$POS^{\mathbb{U}_i}_{\mathbb{C}_{i+1}}(\mathbb{D}) = POS^{\mathbb{U}_i}_{\mathbb{C}_i}(\mathbb{D}) \cup POS^{\mathbb{U}_{i+1}}_{\mathcal{A}_{i+1}}(\mathbb{D}); \tag{28}$$

$$\text{where } \mathbb{U}_{i+1} = \mathbb{U}_i - POS^{\mathbb{U}_i}_{\mathbb{C}_i}(\mathbb{D}) = \{x_j | \mathrm{v}_j(\mathbb{C}_i) = 1\}; \tag{29}$$

and $\mathbb{U}_1 = \mathbb{U}$. In effect, the decision attribute set $\mathbb{D}$ can be positively approximated using granulation orders $\mathbb{C}_i$ and $\mathbb{C}_{i+1}$ on the gradually reduced universe, respectively.

The change in dependency when an attribute is removed from the set of condition attributes, is a measure of the significance of the attribute. To what extent an attribute is contributing to calculate the dependency on decision attribute can be calculated by the significance of that attribute. Combining (5), (24), and (27), the significance of the attribute $\mathcal{A}_k$ with respect to the condition attribute set $\mathbb{C}$ is given by

$$\sigma_{\mathbb{C}}(\mathbb{D}, \mathcal{A}_k) = \frac{1}{n} \sum_{j=1}^{n} [\mathrm{v}_j(\mathbb{C} - \{\mathcal{A}_k\}) - \mathrm{v}_j(\mathbb{C})]; \tag{30}$$

where $0 \leq \sigma_{\mathbb{C}}(\mathbb{D}, \mathcal{A}_k) \leq 1$. Hence, the higher the change in dependency, the more significant the attribute $\mathcal{A}_k$ is. If significance is 0, then the attribute is dispensable.

4 Proposed Feature Selection Method

The real life high dimensional data set may contain a 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. The selected feature subset should contain the features those have high relevance and high dependency with the classes 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 paper, rough set theory and hypercuboid equivalence partition matrix are used to select relevant and significant features from high dimensional data sets.

4.1 Rough Hypercuboid for Feature Selection

Let $\mathbb{C} = \{\mathcal{A}_1, \cdots, \mathcal{A}_i, \cdots, \mathcal{A}_j, \cdots, \mathcal{A}_m\}$ denotes the set of m features of a given data set and $\mathbb{S}$ is the set of selected features. Define $\gamma_{\mathcal{A}_i}(\mathbb{D})$ as the relevance of the feature $\mathcal{A}_i$ with respect to the class labels $\mathbb{D}$, $\gamma_{\mathbb{S}}(\mathbb{D})$ as the dependency of the class labels $\mathbb{D}$ on the selected feature set $\mathbb{S}$, while $\sigma_{\{\mathcal{A}_i, \mathcal{A}_j\}}(\mathbb{D}, \mathcal{A}_i)$ as the significance of the feature $\mathcal{A}_i$ with respect to another feature $\mathcal{A}_j \in \mathbb{S}$.

The average relevance of all selected features is, therefore, given by

$$\mathcal{J}_{\text{relev}} = \frac{1}{|\mathbb{S}|} \sum_{\mathcal{A}_i \in \mathbb{S}} \gamma_{\mathcal{A}_i}(\mathbb{D}), \tag{31}$$

the dependency of the class labels $\mathbb{D}$ on the selected feature set $\mathbb{S}$ is given by

$$\mathcal{J}_{\text{depen}} = \gamma_{\mathbb{S}}(\mathbb{D}), \tag{32}$$

while average significance among selected features is

$$\mathcal{J}_{\text{signf}} = \frac{\sum_{\mathcal{A}_i \neq \mathcal{A}_j \in \mathbb{S}} \{\sigma_{\{\mathcal{A}_i, \mathcal{A}_j\}}(\mathbb{D}, \mathcal{A}_i) + \sigma_{\{\mathcal{A}_i, \mathcal{A}_j\}}(\mathbb{D}, \mathcal{A}_j)\}}{|\mathbb{S}|(|\mathbb{S}| - 1)} \tag{33}$$

Therefore, the problem of selecting a set $\mathbb{S}$ of relevant and significant features from the whole feature set $\mathbb{C}$ is equivalent to maximize $\mathcal{J}_{\text{relev}}$, $\mathcal{J}_{\text{depen}}$, and $\mathcal{J}_{\text{signf}}$, that is, to maximize the objective function $\mathcal{J}$, where

$$\mathcal{J} = \omega \mathcal{J}_{\text{relev}} + (1 - \omega)[\lambda \mathcal{J}_{\text{depen}} + (1 - \lambda)\mathcal{J}_{\text{signf}}] \tag{34}$$

where ω and λ are two weight parameters. To solve the above problem, the following greedy algorithm is used.

1) Initialize $\mathbb{C} \leftarrow \{\mathcal{A}_1, \cdots, \mathcal{A}_i, \cdots, \mathcal{A}_j, \cdots, \mathcal{A}_m\}, \mathbb{S} \leftarrow \emptyset$.
2) Calculate the relevance $\gamma_{\mathcal{A}_i}(\mathbb{D})$ of feature $\mathcal{A}_i \in \mathbb{C}$.
3) Select $\mathcal{A}_i$ as most relevant feature that has highest relevance $\gamma_{\mathcal{A}_i}(\mathbb{D})$. In effect, $\mathcal{A}_i \in \mathbb{S}$ and $\mathbb{C} = \mathbb{C} \setminus \mathcal{A}_i$.
4) Repeat the following two steps until $\mathbb{C} = \emptyset$ (or the desired number of features d is selected).
5) Repeat the following two steps for each of the remaining features of $\mathbb{C}$.
   a) Calculate dependency and average significance of each feature $\mathcal{A}_j \in \mathbb{C}$ with respect to the already selected features of $\mathbb{S}$.
   b) Remove $\mathcal{A}_j$ from $\mathbb{C}$ if the dependency remains same with respect to the already selected features. In effect, $\mathbb{C} = \mathbb{C} \setminus \mathcal{A}_j$.
6) From the remaining features of $\mathbb{C}$, select feature $\mathcal{A}_j$ that maximizes the following condition:

$$\omega \gamma_{\mathcal{A}_j}(\mathbb{D}) + \lambda(1 - \omega)[\gamma_{\{\mathbb{S}, \mathcal{A}_j\}}(\mathbb{D}) - \gamma_{\mathbb{S}}(\mathbb{D})] + \frac{(1 - \omega)(1 - \lambda)}{|\mathbb{S}|} \sum_{\mathcal{A}_i \in \mathbb{S}} \sigma_{\{\mathcal{A}_i, \mathcal{A}_j\}}(\mathbb{D}, \mathcal{A}_j). \tag{35}$$

As a result of that, $\mathcal{A}_j \in \mathbb{S}$ and $\mathbb{C} = \mathbb{C} \setminus \mathcal{A}_j$.

Both relevance $\gamma_{\mathcal{A}_i}(\mathbb{D})$ of a feature $\mathcal{A}_i$ with respect to class labels $\mathbb{D}$ and dependency $\gamma_{\mathbb{S}}(\mathbb{D})$ of class labels $\mathbb{D}$ on the selected feature set $\mathbb{S}$ are calculated using (24), while significance $\sigma_{\mathbb{S}}(\mathbb{D}, \mathcal{A}_i)$ of the feature $\mathcal{A}_i$ with respect to the set $\mathbb{S}$ is computed using (30).

6 Experimental Results

The performance of proposed hypercuboid equivalence partition matrix based feature selection method is extensively studied and compared with that of some existing feature selection algorithms. The performance of "Max-Relevance Max-Dependency Max-Significance" (MRMDMS) criterion is also compared with that of different combinations of the individual components of the proposed hybrid criterion. All the algorithms are implemented in C language and run in Ubuntu 10.04 having machine configuration Pentium D, 2.66 GHz, 2 MB cache, and 4 GB RAM. Several benchmark data sets, namely, Iris, Satimage, Isolet, Ionosphere, Segmentation and Multiple Features of *UCI Machine Learning Repository* [31], and Breast Cancer, Lung Cancer, DLBCLNIH, Prostate Cancer and Leukemia of *Kent Ridge Bio-Medical Data Set Repository* [32] are used to evaluate the performance of different methods. The leave-one-out cross-validation is performed on Isolet, Ionosphere and Multiple Features data sets, while training-testing is done on remaining data sets.

6.1 Feature Evaluation Indices

Following five feature evaluation indices are used to evaluate the performance of different algorithms.

6.1.1 Class Separability Index

The class separability index $\mathcal{S}$ [2] of a data set is defined as $\mathcal{S} = \text{trace}(V_B^{-1}V_W)$, where V_W is the within class scatter matrix and V_B is the between class scatter matrix, defined as follows:

$$V_W = \sum_{j=1}^{c} \pi_j E\{(X-\mu_j)(X-\mu_j)^T|\beta_j\} = \sum_{j=1}^{c} \pi_j \Sigma_j;$$

$$V_B = \sum_{j=1}^{c} \pi_j (\mu_j - \bar{\mu})(\mu_j - \bar{\mu})^T; \text{and } \bar{\mu} = E\{X\} = \sum_{j=1}^{c} \pi_j \mu_j;$$

where c is the number of classes, π_j is a priori probability that a pattern belongs to class β_j, X is a feature vector, $\bar{\mu}$ is the sample mean vector for the entire data points, μ_j and Σ_j represent the sample mean and covariance matrix of class β_j, respectively, and $E\{\cdot\}$ is the expectation operator. A lower value of $\mathcal{S}$ ensures that classes are well separated by their scatter means.

6.1.2 Entropy

The entropy E of a data set is defined as [33]:

$$E = -\sum_{i=1}^{n}\sum_{j=1}^{n} \mathrm{S}_{ij} \times \log \mathrm{S}_{ij} + (1-\mathrm{S}_{ij}) \times \log(1-\mathrm{S}_{ij}) \quad (42)$$

where $\mathrm{S}_{ij} = e^{-\alpha \mathrm{D}_{ij}}$ represents the similarity between two objects x_i and x_j, α is $\frac{-\ln 0.5}{\bar{\mathrm{D}}}$, $\bar{\mathrm{D}}$ is the average distance between data points computed over the entire data set, and the distance D_{ij} between x_i and x_j is

$$\mathrm{D}_{ij} = \sqrt{\left[\sum_{k=1}^{d}\left(\frac{x_{ik}-x_{jk}}{\max_k - \min_k}\right)^2\right]} \quad (43)$$

where x_{ik} is the feature value for x_i along kth axis, $\max_k$, $\min_k$ are the maximum and minimum values computed over all samples along kth axis, and d is the number of features. If the data is uniformly distributed in feature space, entropy is maximum. When the data has well-formed clusters uncertainty is low and so is entropy.

6.1.3 Representation Entropy

The representation entropy is defined as [2]

$$H_R = -\sum_{j=1}^{d} \tilde{\lambda}_j \log \tilde{\lambda}_j; \text{ where } \tilde{\lambda}_j = \frac{\lambda_j}{\sum_{j=1}^{d}\lambda_j} \quad (44)$$

and $\lambda_j, j = 1, \cdots, d$ are the eigenvalues of $d \times d$ covariance matrix of a feature set of size d. The function H_R attains a minimum value zero when all eigenvalues except one are zero, that is, when all the information is present along a single coordinate direction. If all the eigenvalues are equal, that is, information is equally distributed among all the features, H_R is maximum and so is the uncertainty involved in feature reduction.

6.1.4 Support Vector Machine

The support vector machine (SVM) [34] is a margin classifier that draws an optimal hyperplane in the feature vector space; this defines a boundary that maximizes the margin between samples in different classes, therefore leading to good generalization properties. A key factor in the SVM is to use kernels to construct nonlinear decision boundary. In the present work, linear kernels are used.

6.1.5 C4.5 Decision Tree

The C4.5 [35] is a decision tree based classification algorithm. It is used for evaluating the effectiveness of reduced feature set for classification. It performs feature selection in the process of training and the classification models it builds are represented in the form of decision trees, which can be further examined.

6.2 Result on Iris Data

The relevance values of four features of Iris data obtained using the proposed feature selection method are 0.1867, 0.0467, 0.7533, and 0.7467, respectively, considering $\omega = \lambda = 0.5$. Hence, the Feature 3 is selected first as it has highest relevance value. After selecting Feature 3, the dependency, overall significance, and average significance of each feature are calculated.

Dependency of each feature:
Feature 1: 0.7533; Feature 2: 0.7533; Feature 4: 0.8267
Overall significance of each feature:
Feature 1: 0.0000; Feature 2: 0.0000; Feature 4: 0.0733
Average significance of each feature:
Feature 1: 0.0000; Feature 2: 0.0000; Feature 4: 0.0733

Based on the values of overall and average significance, two features, namely, Features 1 and 2, are eliminated as they are insignificant features with respect to

TABLE 1
Performance of Proposed Algorithm on Different Data Sets ($\omega = 0.5$, $\lambda = 0.5$)

Different Data Sets	Selected Features	Performance on Training Set					Performance on Test Set				
		SVM	C4.5	α index	$\alpha^\star$ index	γ index	SVM	C4.5	α index	$\alpha^\star$ index	γ index
Breast Cancer $m = 24481$, $c = 2$ $n = 78 : 19$	1	57.69	74.35	0.163	0.164	0.282	36.84	57.89	0.400	0.407	0.579
	2	69.23	82.05	0.321	0.322	0.487	52.63	57.89	0.458	0.462	0.632
	3	70.51	89.74	0.429	0.431	0.603	68.42	63.15	0.571	0.583	0.737
	4	70.51	89.74	0.520	0.529	0.692	73.68	63.15	0.635	0.652	0.789
	5	70.51	89.74	0.586	0.592	0.744	73.68	63.15	0.705	0.727	0.842
	6	70.51	92.30	0.686	0.696	0.821	73.68	68.42	0.786	0.810	0.895
	7	75.64	92.30	0.727	0.733	0.846	78.95	73.68	0.786	0.810	0.895
Satimage $m = 36$, $c = 6$ $n = 4435 : 2000$	1	59.14	61.12	0.026	0.021	0.084	57.15	58.45	0.025	0.021	0.091
	2	78.47	82.29	0.053	0.036	0.124	77.10	77.15	0.060	0.038	0.133
	3	80.16	85.34	0.057	0.038	0.129	77.65	78.45	0.062	0.040	0.135
	4	80.74	87.62	0.059	0.041	0.133	77.85	78.45	0.063	0.042	0.139
	5	81.58	90.30	0.061	0.042	0.137	78.85	79.40	0.067	0.046	0.149
	6	81.69	91.13	0.062	0.043	0.139	79.10	79.90	0.068	0.047	0.150
	7	86.25	93.34	0.072	0.057	0.170	83.90	83.50	0.076	0.061	0.183
	8	86.38	94.11	0.073	0.058	0.173	84.25	83.75	0.078	0.063	0.186
	9	86.56	94.74	0.074	0.059	0.174	84.45	83.75	0.079	0.065	0.188
	10	86.65	95.42	0.095	0.084	0.231	84.65	84.30	0.144	0.133	0.331
Segmentation $m = 18$, $c = 7$ $n = 210 : 2100$	1	48.09	61.90	0.090	0.055	0.200	44.90	42.85	0.005	0.003	0.014
	2	72.86	88.57	0.312	0.203	0.414	70.67	74.38	0.067	0.052	0.166
	3	78.09	90.48	0.419	0.277	0.481	76.52	76.23	0.228	0.121	0.295
	4	88.57	96.19	0.626	0.525	0.710	88.86	88.42	0.435	0.268	0.476
	5	91.43	97.14	0.679	0.606	0.776	90.38	89.33	0.462	0.321	0.523
	6	92.38	97.14	0.687	0.618	0.786	90.86	89.33	0.465	0.325	0.525
	7	93.33	97.14	0.729	0.684	0.833	92.76	89.33	0.468	0.328	0.527
	8	93.81	97.14	0.739	0.697	0.843	92.62	89.33	0.469	0.329	0.528
	9	93.81	97.14	0.752	0.715	0.848	92.67	89.62	0.470	0.330	0.529
	10	93.81	97.14	0.752	0.715	0.848	95.24	90.00	0.472	0.332	0.529

the already selected feature, namely, Feature 3; only the remaining feature, namely, Feature 4 is selected next as the second feature that has the objective function value 0.41. Fig. 2 presents the scatter plots of samples from three classes of Iris data set, along with three two-dimensional class hypercuboids constructed with Features 3 and 4. Each hypercuboid encloses a region in the two-dimensional space. However, the class hypercuboids for Classes 2 and 3 intersect with each other and the intersection forms an implicit hypercuboid. The implicit hypercuboid encompasses the misclassified samples or objects those belong to more than one classes. The values of different quantitative indices for Features 3 and 4 are reported next, along with that for whole feature set.

Measures/Features	3 and 4	1 to 4
Classification accuracy, SVM	95.33%	96.00%
Classification accuracy, C4.5	98.00%	98.00%
Class separability index, $\mathcal{S}$	0.069	0.389
Entropy, E	0.690	0.741
Representation entropy, H_R	0.997	0.879
Average accuracy, α	0.723	0.723
Accuracy of approximation, α^*	0.705	0.705
Quality of approximation, γ	0.827	0.827

The results reported above establish that the proposed method selects most effective features from the whole feature set by maximizing relevance, dependency, and significance of the selected features.

6.3 Effectiveness of Proposed Method

To better understand the effectiveness of the proposed method, extensive experimental results are reported in

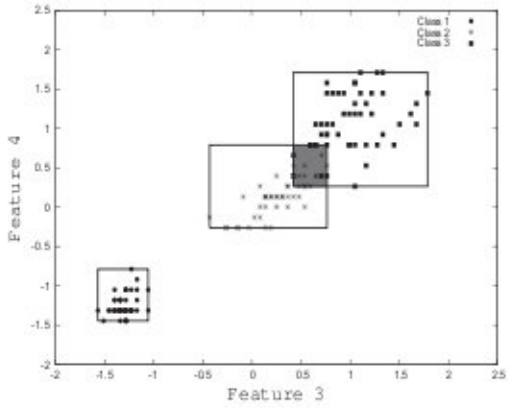

Fig. 2. Rough hypercuboids for Iris data in two dimension

Table 1 for Breast Cancer, Satimage, and Segmentation data sets. Subsequent discussions analyze the results with respect to various proposed quantitative indices such as α, $\alpha^\star$, and γ, and the classification accuracy of both SVM and C4.5. Results are presented for different values of the number of selected features d, considering weight parameters $\omega = \lambda = 0.5$. All the results reported here confirm that as the number of selected features d increases, the classification accuracy of both SVM and C4.5 as well as the values of $\alpha, \alpha^\star$, and γ indices increase. Finally, all the indices are saturated when d equals to the number of selected features for each data set. Hence, the proposed indices such as α, $\alpha^\star$, and γ can be used to act as the objective function of feature selection algorithm in approximation spaces as they reflect good quantitative measures like existing SVM and C4.5. Also, the values of different indices for training and testing confirm that the proposed feature selection algorithm

TABLE 2
Performance of Proposed Feature Evaluation Criterion on Different Data Sets

Different Data Sets	Selected Features	Parameters		Different Evaluation Criteria							
		ω	λ	SVM (%)	C4.5 (%)	$\mathcal{S}$ index	E	H_R	α index	$\alpha^\star$ index	γ index
Breast Cancer $m = 24481$ $n = 78 : 19$ $c = 2$	$d = 7$	1.0	*	68.42	63.15	9.859	0.779	2.653	0.778	0.810	0.895
		0.0	0.0	63.16	57.89	12.621	0.781	2.697	0.786	0.810	0.895
			1.0	52.63	68.42	15.714	0.779	2.676	0.890	0.900	0.947
			0.5	63.16	57.89	9.706	0.782	2.585	0.786	0.800	0.847
		0.5	0.0	78.95	73.68	9.879	0.777	2.661	0.786	0.810	0.895
			1.0	78.95	73.68	9.879	0.777	2.661	0.786	0.810	0.895
			0.5	78.95	73.68	9.879	0.777	2.661	0.786	0.810	0.895
Satimage $m = 36$ $n = 4435 : 2000$ $c = 6$	$d = 10$	1.0	*	83.20	83.10	0.411	0.755	3.223	0.084	0.069	0.204
		0.0	0.0	84.50	83.70	0.372	0.758	3.215	0.087	0.077	0.220
			1.0	84.65	85.60	0.416	0.767	3.193	0.149	0.135	0.338
			0.5	85.15	85.20	0.385	0.760	3.220	0.143	0.129	0.326
		0.5	0.0	84.80	82.95	0.383	0.755	3.244	0.082	0.070	0.200
			1.0	84.65	84.30	0.372	0.754	3.243	0.144	0.133	0.331
			0.5	84.65	84.30	0.372	0.754	3.243	0.144	0.133	0.331
Isolet $m = 617$ $n = 7797$ $c = 26$	$d = 25$	1.0	*	58.27	87.64	1.679	0.276	4.416	0.007	0.004	0.044
		0.0	0.0	60.31	86.91	1.517	0.276	4.474	0.017	0.005	0.057
			1.0	84.67	94.18	0.919	0.276	4.433	0.101	0.055	0.258
			0.5	86.89	94.63	0.815	0.276	4.416	0.009	0.058	0.260
		0.5	0.0	57.01	87.71	1.595	0.276	4.468	0.008	0.003	0.044
			1.0	58.41	94.88	0.789	0.276	4.464	0.093	0.058	0.257
			0.5	88.92	95.20	0.782	0.276	4.478	0.101	0.072	0.265
Leukemia $m = 12558$ $n = 215 : 112$ $c = 7$	$d = 49$	1.0	*	75.89	68.04	0.788	0.793	3.518	0.949	0.948	0.973
		0.0	0.0	81.25	78.57	0.865	0.793	3.524	0.960	0.965	0.982
			1.0	80.36	77.68	1.800	0.810	3.297	0.923	0.931	0.964
			0.5	83.04	76.79	0.841	0.798	3.309	0.932	0.948	0.973
		0.5	0.0	81.25	75.00	0.866	0.791	3.523	0.960	0.965	0.982
			1.0	84.82	71.43	0.844	0.792	3.652	0.943	0.965	0.982
			0.5	89.29	80.36	0.716	0.763	3.711	0.960	0.965	0.982
Ionosphere $m = 33$ $n = 351$ $c = 2$	$d = 9$	1.0	*	86.32	91.73	9.168	0.759	3.126	0.162	0.194	0.325
		0.0	0.0	86.32	91.73	8.897	0.758	3.126	0.162	0.194	0.325
			1.0	87.18	95.72	9.168	0.759	3.126	0.164	0.196	0.328
			0.5	87.18	95.72	9.168	0.759	3.126	0.164	0.196	0.328
		0.5	0.0	86.32	95.72	9.168	0.759	3.126	0.162	0.194	0.325
			1.0	87.18	94.58	9.168	0.759	3.126	0.164	0.196	0.328
			0.5	87.75	98.00	8.897	0.758	3.126	0.164	0.196	0.328
Segmentation $m = 18$ $n = 210 : 2100$ $c = 7$	$d = 10$	1.0	*	91.05	90.00	0.747	0.775	2.271	0.470	0.329	0.528
		0.0	0.0	92.19	89.38	0.521	0.769	2.271	0.475	0.333	0.531
			1.0	92.86	89.29	0.789	0.793	2.006	0.472	0.332	0.522
			0.5	95.24	90.00	0.681	0.793	1.977	0.472	0.329	0.528
		0.5	0.0	92.67	89.67	0.512	0.769	2.275	0.471	0.330	0.530
			1.0	92.24	89.62	0.699	0.769	2.400	0.470	0.322	0.520
			0.5	95.24	90.00	0.159	0.769	2.400	0.472	0.332	0.529

can generalize a data set irrespective of the number of original and selected features, classes, and samples. However, the classification accuracy of the SVM using 7 selected features on test set of Segmentation data is higher than that obtained using 8 and 9 features.

6.4 Importance of Relevance and Significance

To establish the effectiveness of proposed MRMDMS criterion for feature selection over other criteria, extensive experimental results are reported in Table 2 with respect to classification accuracy of both SVM and C4.5, and different quantitative indices such as $\mathcal{S}$, H_R, E, $\alpha, \alpha^\star$, and γ. The results obtained using the MR, MD, MS, MRMD, MRMS, and MDMS criteria, which are equivalent to the MRMDMS criterion with $\omega = 1.0$, $\{\omega = 0.0, \lambda = 1.0\}$, $\{\omega = 0.0, \lambda = 0.0\}$, $\{\omega = 0.5, \lambda = 1.0\}$, $\{\omega = 0.5, \lambda = 0.0\}$, and $\{\omega = 0.0, \lambda = 0.5\}$, respectively, are also presented in this table for the sake of comparison,

From the results reported in Table 2, it is seen that the performance of proposed MRMDMS criterion is better than that of MR and MRMD criteria irrespective of the data sets and quantitative indices used, and that of other criteria in most of the cases. The MD criterion achieves better performance in terms of α, $\alpha^\star$, and γ indices for Breast Cancer and Satimage data, H_R value for Breast Cancer data, and classification accuracy of the C4.5 for Satimage data. The MS criterion provides higher values of H_R for Breast Cancer data and α, $\alpha^\star$, and γ for Segmentation data. On the other hand, the MRMS criterion attains higher classification accuracy of the SVM and H_R value only for Satimage data, while the MDMS criterion achieves higher classification accuracy of both SVM and C4.5 for Satimage data, and lower value of $\mathcal{S}$ index for Breast Cancer data. Out of total 288 cases, the proposed MRMDMS criterion achieves significantly better results than other criteria in 270 cases. Hence, the proposed criterion must be used to get a reduced set of relevant and significant features.

6.5 Optimum Value of Weight Parameter ω

The parameter ω regulates the relative importance of the significance, both average and overall, of the candidate

TABLE 3
Performance on Various Data Sets for Different Values of Weight Parameter ω ($\lambda = 0.5$)

Different Data Sets	Evaluation Criteria	Value of Weight Parameter ω										
		0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Breast Cancer $d = 7$	SVM (%)	63.16	63.16	63.16	73.68	78.95	78.95	78.95	78.95	78.95	78.95	68.42
	C4.5 (%)	57.89	68.42	68.42	68.42	73.68	73.68	73.68	73.68	63.16	63.15	63.15
	$\mathcal{S}$ index	9.706	9.706	9.706	12.761	9.879	9.879	9.879	9.879	9.879	9.879	9.859
	E	0.782	0.782	0.782	0.782	0.777	0.777	0.777	0.777	0.777	0.777	0.779
	H_R	2.585	2.585	2.585	2.591	2.661	2.661	2.661	2.661	2.661	2.661	2.653
	α index	0.786	0.776	0.776	0.776	0.786	0.786	0.786	0.786	0.786	0.786	0.778
	$\alpha^\star$ index	0.800	0.800	0.800	0.800	0.810	0.810	0.810	0.810	0.810	0.810	0.810
	γ index	0.847	0.847	0.847	0.847	0.895	0.895	0.895	0.895	0.895	0.895	0.895
Satimage $d = 10$	SVM (%)	85.15	84.40	84.40	84.65	84.65	84.65	84.20	84.20	83.20	83.20	83.20
	C4.5 (%)	85.20	83.85	83.85	84.30	84.30	84.30	83.45	83.45	83.10	83.10	83.10
	$\mathcal{S}$ index	0.385	0.361	0.361	0.372	0.372	0.372	0.400	0.400	0.411	0.411	0.411
	E	0.760	0.757	0.757	0.754	0.754	0.754	0.759	0.759	0.755	0.755	0.755
	H_R	3.220	3.214	3.214	3.243	3.243	3.243	3.222	3.222	3.223	3.223	3.223
	α index	0.143	0.140	0.140	0.144	0.144	0.144	0.079	0.079	0.084	0.084	0.084
	$\alpha^\star$ index	0.129	0.128	0.128	0.133	0.133	0.133	0.065	0.065	0.069	0.069	0.069
	γ index	0.326	0.323	0.323	0.331	0.331	0.331	0.190	0.190	0.204	0.204	0.204
Isolet $d = 25$	SVM (%)	86.89	86.61	88.41	88.47	88.71	88.92	86.56	78.77	64.78	58.29	58.27
	C4.5 (%)	94.63	94.32	95.31	95.34	94.82	95.20	94.20	92.48	89.46	88.19	87.64
	$\mathcal{S}$ index	0.815	0.813	0.775	0.774	0.817	0.782	0.959	1.293	1.491	1.496	1.679
	E	0.276	0.276	0.276	0.276	0.276	0.276	0.276	0.276	0.276	0.276	0.276
	H_R	4.416	4.418	4.397	4.399	4.409	4.478	4.434	4.347	4.426	4.459	4.416
	α index	0.090	0.089	0.090	0.091	0.090	0.101	0.057	0.030	0.010	0.008	0.007
	$\alpha^\star$ index	0.058	0.055	0.063	0.067	0.063	0.072	0.041	0.021	0.006	0.004	0.004
	γ index	0.260	0.254	0.262	0.260	0.258	0.265	0.202	0.142	0.053	0.045	0.044
Leukemia $d = 49$	SVM (%)	83.04	86.61	88.39	88.39	88.39	89.29	88.39	79.46	81.25	81.25	75.89
	C4.5 (%)	76.79	65.18	67.86	67.86	68.75	80.36	68.75	74.11	71.43	71.43	68.04
	$\mathcal{S}$ index	0.841	0.842	0.825	0.795	0.807	0.716	0.788	0.747	0.725	0.725	0.788
	E	0.798	0.797	0.796	0.795	0.795	0.763	0.793	0.788	0.782	0.782	0.793
	H_R	3.309	3.526	3.715	3.716	3.714	3.711	3.711	3.540	3.539	3.539	3.518
	α index	0.932	0.960	0.960	0.960	0.960	0.960	0.960	0.960	0.951	0.951	0.949
	$\alpha^\star$ index	0.948	0.965	0.965	0.965	0.965	0.965	0.965	0.965	0.962	0.962	0.948
	γ index	0.973	0.982	0.982	0.982	0.982	0.982	0.982	0.982	0.981	0.981	0.973
Ionosphere $d = 9$	SVM (%)	87.18	87.18	87.18	87.18	87.18	87.75	87.18	87.18	87.18	87.18	86.32
	C4.5 (%)	95.72	91.73	91.73	91.73	91.73	98.00	91.73	91.73	91.73	91.73	91.73
	$\mathcal{S}$ index	9.168	9.168	9.168	9.168	9.168	8.897	9.168	9.168	9.168	9.168	9.168
	E	0.759	0.759	0.759	0.759	0.759	0.758	0.759	0.759	0.759	0.759	0.759
	H_R	3.126	3.126	3.126	3.126	3.126	3.126	3.126	3.126	3.126	3.126	3.126
	α index	0.164	0.164	0.164	0.164	0.164	0.164	0.164	0.164	0.164	0.164	0.162
	$\alpha^\star$ index	0.196	0.196	0.196	0.196	0.196	0.196	0.196	0.196	0.196	0.196	0.194
	γ index	0.328	0.328	0.328	0.328	0.328	0.328	0.328	0.328	0.328	0.328	0.325
Segmentation $d = 10$	SVM (%)	95.24	95.24	92.95	92.24	92.24	95.24	92.43	92.43	92.48	92.48	91.05
	C4.5 (%)	90.00	90.00	89.52	89.52	89.52	90.00	89.62	89.62	89.76	89.76	90.00
	$\mathcal{S}$ index	0.681	0.681	0.681	0.700	0.700	0.159	0.974	0.974	0.747	0.747	0.747
	E	0.793	0.793	0.788	0.770	0.770	0.769	0.792	0.792	0.775	0.775	0.775
	H_R	1.977	1.977	2.261	2.386	2.386	2.400	2.328	2.328	2.271	2.271	2.271
	α index	0.472	0.472	0.472	0.471	0.471	0.472	0.470	0.470	0.470	0.470	0.470
	$\alpha^\star$ index	0.329	0.329	0.330	0.332	0.332	0.332	0.332	0.332	0.332	0.332	0.329
	γ index	0.528	0.528	0.522	0.529	0.529	0.529	0.522	0.522	0.521	0.521	0.528

feature with respect to the already selected features and the relevance with the output class. If ω is one, only the relevance of each feature with the output class is considered. If the significance between features is not taken into account, selecting the features with the highest relevance with respect to the output class may tend to produce a set of redundant and insignificant features that may leave out useful complementary information. On the other hand, if ω is zero, the features are selected based on their average and overall significance values only without considering the relevance of each feature. In effect, the selected feature set may contain a number of irrelevant features. Hence, the value of weight parameter ω should be in between zero and one in order to obtain good results, that is, $0 < \omega < 1$.

Table 3 presents the performance of proposed feature selection method for different values of ω considering $\lambda = 0.5$. From the results reported in Table 3, it is seen that as the value of ω increases, the classification accuracy of both SVM and C4.5, and the values of H_R, α, $\alpha^\star$, and γ indices increase, whereas the values of $\mathcal{S}$ index and E decrease. The performance of the proposed method detoriates when $\omega = 0.0$ and 1.0 as the selected feature set may contain irrelevant features for $\omega = 0.0$ and redundant and insignificant features for $\omega = 1.0$. The proposed method achieves its best performance at $0.4 \leq \omega \leq 0.7$ for Breast Cancer data, $0.3 \leq \omega \leq 0.5$ for Satimage data, and $\omega = 0.5$ for Isolet, Leukemia, Ionosphere, and Segmentation data with respect to most of the quantitative indices. Hence, it provides its best performance for $\omega = 0.5$ irrespective of the quantitative indices and data sets used. In other words, the best performance of the proposed method is achieved when nearly equal weightage is given to both relevance

TABLE 4
Performance on Various Data Sets for Different Values of Weight Parameter λ ($\omega = 0.5$)

Different Data Sets	Evaluation Criteria	Value of Weight Parameter λ										
		0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Satimage $d = 10$	SVM (%)	84.80	84.80	84.80	84.80	84.80	84.65	84.65	84.65	84.65	84.65	84.65
	C4.5 (%)	82.95	82.95	82.95	82.95	82.95	84.30	84.30	84.30	84.30	84.30	84.30
	$\mathcal{S}$ index	0.383	0.383	0.383	0.383	0.383	0.372	0.372	0.372	0.372	0.372	0.372
	E	0.755	0.755	0.755	0.755	0.755	0.754	0.754	0.754	0.754	0.754	0.754
	H_R	3.244	3.244	3.244	3.244	3.244	3.243	3.243	3.243	3.243	3.243	3.243
	α index	0.082	0.082	0.082	0.082	0.082	0.144	0.144	0.144	0.144	0.144	0.144
	$\alpha^\star$ index	0.070	0.070	0.070	0.070	0.070	0.133	0.133	0.133	0.133	0.133	0.133
	γ index	0.200	0.200	0.200	0.200	0.200	0.331	0.331	0.331	0.331	0.331	0.331
Isolet $d = 25$	SVM (%)	57.01	75.34	85.33	87.33	87.23	88.92	88.71	88.69	87.91	87.98	58.41
	C4.5 (%)	87.71	87.65	91.84	94.49	94.46	95.20	95.10	94.82	95.01	94.77	94.88
	$\mathcal{S}$ index	1.595	1.476	1.260	0.907	0.903	0.782	0.880	0.817	0.781	0.807	0.789
	E	0.276	0.276	0.276	0.276	0.276	0.276	0.276	0.276	0.276	0.276	0.276
	H_R	4.468	4.473	4.339	4.439	4.434	4.478	4.394	4.409	4.426	4.466	4.464
	α index	0.008	0.007	0.034	0.058	0.061	0.101	0.081	0.087	0.090	0.096	0.093
	$\alpha^\star$ index	0.003	0.004	0.018	0.041	0.044	0.072	0.063	0.063	0.065	0.058	0.058
	γ index	0.044	0.044	0.133	0.199	0.210	0.265	0.254	0.258	0.262	0.261	0.257
Leukemia $d = 49$	SVM (%)	81.25	85.71	88.39	88.39	88.39	89.29	88.39	88.39	89.29	87.50	84.82
	C4.5 (%)	75.00	75.00	68.75	68.75	68.75	80.36	68.75	68.75	68.75	72.32	71.43
	$\mathcal{S}$ index	0.866	0.761	0.805	0.803	0.788	0.716	0.788	0.788	0.787	0.798	0.844
	E	0.791	0.792	0.793	0.794	0.793	0.763	0.793	0.793	0.793	0.794	0.792
	H_R	3.523	3.512	3.690	3.703	3.711	3.711	3.711	3.711	3.709	3.648	3.652
	α index	0.960	0.960	0.960	0.960	0.960	0.960	0.960	0.960	0.960	0.960	0.943
	$\alpha^\star$ index	0.965	0.965	0.965	0.965	0.965	0.965	0.965	0.965	0.965	0.965	0.965
	γ index	0.982	0.982	0.982	0.982	0.982	0.982	0.982	0.982	0.982	0.982	0.982
Ionosphere $d = 9$	SVM (%)	86.32	87.18	87.18	87.18	87.18	87.75	87.18	87.18	87.18	87.18	87.18
	C4.5 (%)	95.72	95.72	95.72	98.00	98.00	98.00	98.00	94.58	94.58	94.58	94.58
	$\mathcal{S}$ index	9.168	9.168	9.168	9.168	9.168	8.897	9.168	9.168	9.168	9.168	9.168
	E	0.759	0.759	0.759	0.759	0.759	0.758	0.759	0.759	0.759	0.759	0.759
	H_R	3.126	3.126	3.126	3.126	3.126	3.126	3.126	3.126	3.126	3.126	3.126
	α index	0.162	0.164	0.164	0.164	0.164	0.164	0.164	0.164	0.164	0.164	0.164
	$\alpha^\star$ index	0.194	0.196	0.196	0.196	0.196	0.196	0.196	0.196	0.196	0.196	0.196
	γ index	0.325	0.328	0.328	0.328	0.328	0.328	0.328	0.328	0.328	0.328	0.328
Segmentation $d = 10$	SVM (%)	92.67	92.43	92.43	92.24	92.24	95.24	92.24	92.24	92.24	92.24	92.24
	C4.5 (%)	89.67	89.62	89.62	89.62	89.62	90.00	89.62	89.62	89.62	89.62	89.62
	$\mathcal{S}$ index	0.512	0.974	0.974	0.699	0.699	0.159	0.699	0.699	0.699	0.699	0.699
	E	0.769	0.792	0.792	0.769	0.769	0.769	0.769	0.769	0.769	0.769	0.769
	H_R	2.275	2.328	2.328	2.400	2.400	2.400	2.400	2.400	2.400	2.400	2.400
	α index	0.471	0.471	0.471	0.471	0.472	0.472	0.472	0.472	0.470	0.470	0.470
	$\alpha^\star$ index	0.330	0.330	0.330	0.332	0.332	0.332	0.322	0.322	0.322	0.322	0.322
	γ index	0.530	0.530	0.522	0.529	0.529	0.529	0.529	0.529	0.529	0.529	0.520

and significance of each feature selection. However, the values of $\mathcal{S}$ index at $0.0 \leq \omega \leq 0.2$ and $\omega = 1.0$ for Breast Cancer data and $\omega = 0.1$ and 0.2 for Satimage data are lower than that obtained at $\omega = 0.5$. Also, the proposed method provides highest accuracy of both SVM and C4.5 at $\omega = 0.0$ for Satimage data, that of the C4.5 at $\omega = 0.2$ and 0.3 for Isolet data, and highest value of H_R at $0.2 \leq \omega \leq 0.4$ for Leukemia data.

6.6 Optimum Value of Weight Parameter λ

The weight parameter λ controls the relative importance of average significance of the candidate feature with respect to the already selected features and the overall significance of the feature or dependency of the output class on the set of selected features. If λ becomes one, only the dependency of the output class on the selected feature set is considered. On the other hand, the features are selected based on their average significance values only when λ is zero. In effect, the dependency between data distribution in multidimensional space and output class is not considered in this case. Hence, the value of parameter λ should also be in between zero and one in order to obtain better performance, that is, $0 < \lambda < 1$.

Table 4 presents the performance of proposed method for different values of λ considering $\omega = 0.5$. From the results reported in Table 4, it is seen that the proposed method achieves its best performance at $0.5 \leq \lambda \leq 1.0$ for Satimage data, and $\lambda = 0.5$ for Isolet, Leukemia, Ionosphere, and Segmentation data with respect to most of the quantitative indices, while the performance remains unchanged on Breast Cancer data irrespective of the values of λ. That is, it provides its best performance for $\lambda = 0.5$ irrespective of the quantitative indices and data sets used. In other words, the best performance of the proposed method is achieved when nearly equal weightage is given to both average significance and overall significance or dependency for each feature selection. However, the proposed method provides highest classification accuracy of the SVM at $0.0 \leq \lambda \leq 0.4$ for Satimage data, and $0.2 \leq \lambda \leq 0.4$, $\lambda = 0.6$ and 0.7 for Leukemia data. Also, the highest value of γ index for Segmentation data is obtained at $\lambda = 0.0$ and 0.1.

6.7 Performance of Different Rough Set Models

Furthermore, extensive experiments are done to evaluate the performance of different rough set models such as

TABLE 5
Comparative Performance of Various Rough Set Models on Different Data Sets

Different Data Sets	Different Rough Set Models	Different Evaluation Criteria									
		d	SVM (%)	C4.5 (%)	$\mathcal{S}$	E	H_R	α	α^*	γ	Time (ms)
Breast Cancer	Pawlak's Rough Sets	8	63.16	57.90	12.615	0.781	2.497	0.770	0.795	0.785	312744
	Neighborhood Rough Sets	13	73.68	57.90	7.401	0.778	2.661	0.781	0.795	0.866	2163081
	Fuzzy Rough Sets	10	68.42	57.90	8.859	0.778	2.653	0.775	0.800	0.890	1090860
	Rough Hypercuboid	7	78.95	73.68	9.879	0.777	2.661	0.786	0.810	0.895	1344
Satimage	Pawlak's Rough Sets	6	75.19	78.40	0.467	0.754	2.821	0.081	0.104	0.095	78031
	Neighborhood Rough Sets	15	78.40	81.40	0.478	0.754	2.985	0.112	0.126	0.117	95217
	Fuzzy Rough Sets	11	78.40	81.90	0.460	0.755	3.006	0.136	0.130	0.199	91363
	Rough Hypercuboid	10	84.65	84.30	0.372	0.754	3.243	0.144	0.133	0.331	462
Isolet	Pawlak's Rough Sets	19	59.65	84.80	1.679	0.517	4.433	0.080	0.035	0.104	80814755
	Neighborhood Rough Sets	32	78.77	88.60	1.239	0.322	4.464	0.099	0.065	0.217	91488501
	Fuzzy Rough Sets	16	86.56	88.10	1.272	0.279	4.464	0.099	0.069	0.206	98327762
	Rough Hypercuboid	25	88.92	95.20	0.782	0.276	4.478	0.101	0.072	0.265	107101
Leukemia	Pawlak's Rough Sets	22	75.89	67.86	0.866	0.793	3.523	0.949	0.961	0.933	119731822
	Neighborhood Rough Sets	37	78.57	78.57	0.865	0.763	3.625	0.949	0.965	0.970	471501988
	Fuzzy Rough Sets	63	80.36	78.57	0.865	0.763	3.718	0.949	0.960	0.982	531395284
	Rough Hypercuboid	49	89.29	80.36	0.716	0.763	3.711	0.960	0.965	0.982	14654
Ionosphere	Pawlak's Rough Sets	10	81.77	88.30	8.897	0.758	3.126	0.119	0.175	0.237	2816
	Neighborhood Rough Sets	16	82.05	91.73	9.133	0.758	3.126	0.162	0.176	0.311	32174
	Fuzzy Rough Sets	16	82.05	91.73	11.050	0.758	3.126	0.160	0.177	0.309	74199
	Rough Hypercuboid	9	87.75	98.00	8.897	0.758	3.126	0.164	0.196	0.328	11
Segmentation	Pawlak's Rough Sets	8	91.05	89.29	0.521	0.769	2.275	0.474	0.329	0.521	161183
	Neighborhood Rough Sets	14	92.90	89.38	0.512	0.769	2.271	0.474	0.329	0.525	273961
	Fuzzy Rough Sets	11	92.24	89.29	0.512	0.769	2.271	0.471	0.329	0.528	240925
	Rough Hypercuboid	10	95.24	90.00	0.159	0.769	2.400	0.472	0.332	0.529	11
Lung Cancer	Pawlak's Rough Sets	14	73.83	74.50	3.714	0.712	2.628	0.801	0.812	0.817	49517
	Neighborhood Rough Sets	27	81.21	81.21	2.088	0.787	2.904	0.825	0.825	0.833	610648
	Fuzzy Rough Sets	21	79.87	81.88	1.995	0.781	3.118	0.861	0.869	0.890	283829
	Rough Hypercuboid	13	87.92	85.91	1.386	0.750	3.700	1.000	1.000	1.000	157
DLBCLNIH	Pawlak's Rough Sets	24	62.50	62.50	21.649	0.782	2.839	0.276	0.288	0.473	52837
	Neighborhood Rough Sets	52	56.25	62.50	7.392	0.830	3.007	0.276	0.322	0.472	929375
	Fuzzy Rough Sets	77	68.75	65.00	10.895	0.806	3.816	0.309	0.391	0.497	1580148
	Rough Hypercuboid	19	70.00	68.75	19.380	0.810	4.248	0.395	0.416	0.588	651
Prostate Cancer	Pawlak's Rough Sets	6	82.35	82.35	9.028	0.810	1.672	0.813	0.854	0.857	54927
	Neighborhood Rough Sets	18	88.24	91.18	4.285	0.786	1.909	0.902	0.907	0.922	629837
	Fuzzy Rough Sets	11	91.18	94.12	1.994	0.780	2.104	0.971	0.985	1.000	539281
	Rough Hypercuboid	5	94.12	94.12	2.015	0.774	2.322	1.000	1.000	1.000	272
Multiple Features	Pawlak's Rough Sets	17	81.70	79.62	6.026	0.763	1.733	0.561	0.577	0.628	47210
	Neighborhood Rough Sets	52	83.55	82.89	3.599	0.790	1.984	0.574	0.606	0.611	857421
	Fuzzy Rough Sets	31	83.55	83.07	7.691	0.747	1.871	0.599	0.641	0.650	735871
	Rough Hypercuboid	18	90.68	88.21	5.672	0.758	2.806	0.907	0.923	0.928	461

Pawlak's rough sets [3], neighborhood rough sets [23], fuzzy-rough sets [12], and rough hypercuboid approach. Table 5 compares the performance of feature selection using different models considering $\omega = 0.5$ and $\lambda = 0.5$. The best results obtained using Pawlak's rough sets, neighborhood rough sets, and fuzzy-rough sets are presented in this table for the sake of comparison.

From the results reported in Table 5, it is seen that the proposed rough hypercuboid equivalence partition matrix based approach can potentially yield significantly better results compared to other rough set models, irrespective of the data sets and quantitative indices used. However, fuzzy-rough sets provide lower $\mathcal{S}$ index for Breast Cancer, DLBCLNIH and Prostate Cancer data, lower E for DLBCLNIH and Multiple Features data, and higher H_R index for Leukemia data, than the proposed approach. Also, neighborhood rough sets attain lowest $\mathcal{S}$ index for Breast Cancer, DLBCLNIH and Multiple Features data, while Pawlak's rough sets achieve lowest E for Lung Cancer and DLBCLNIH data. Both neighborhood and Pawlak's rough sets attain highest α index for Segmentation data. The better performance of the proposed approach is achieved due to the fact that the hypercuboid equivalence partition matrix evaluates the quality of a feature set through supervised granulation process that utilizes class information of samples. From the results reported in Table 5, it can also be seen that the execution time of the proposed model is significantly lower than that of other models, irrespective of data sets used. The lower execution time of the proposed model is achieved due to its low computational complexity to compute the relevance, significance, and dependency with respect to the number of selected features, total number of features and samples in original data set.

6.8 Performance of Different Algorithms

Finally, Table 6 compares the performance of proposed algorithm with that of various feature selection algorithms such as mutual information based approaches: InfoGain [26] and mRMR framework [29]; rough set based approaches: quick reduct (RSQR) [7], discernibility matrix using genetic algorithm (GADM) [5], [9] and MRMS framework (RSMRMS) [25]; fuzzy-rough set based approaches: quick reduct (FRQR) [17] and mRMR framework (FRmRMR) [20]; and margin based approaches: rel-

TABLE 6
Comparative Classification Accuracy of Different Feature Selection Methods

Different Data Sets	Different Criteria	Mutual Information		Rough Sets/Fuzzy-Rough Sets					Margin Based		Proposed Algorithm
		InfoGain	mRMR	RSQR	GADM	RSMRMS	FRQR	FRmRMR	RELIEF	SIMBA	
Breast	d	40	38	4	6	9	6	26	23	27	7
	SVM (%)	63.16	73.68	63.16	57.89	63.16	52.63	63.16	68.42	68.42	78.95
	C4.5 (%)	68.42	68.42	52.63	57.89	63.16	63.16	63.16	68.42	68.42	73.68
Satimage	d	36	36	5	8	12	7	36	33	34	10
	SVM (%)	86.00	86.00	60.24	81.90	78.40	81.90	86.00	84.65	86.00	84.65
	C4.5 (%)	84.90	84.90	75.20	82.15	78.35	78.40	84.80	84.30	84.30	84.30
Isolet	d	89	78	7	8	16	11	54	73	68	25
	SVM (%)	82.94	84.18	60.24	63.49	77.41	62.81	87.81	86.24	86.81	88.92
	C4.5 (%)	87.50	87.70	84.80	76.23	86.71	87.20	91.20	86.23	94.80	95.20
Leukemia	d	46	32	17	9	20	23	45	74	67	49
	SVM (%)	88.39	84.82	71.43	45.54	81.25	75.89	87.50	87.50	88.39	89.29
	C4.5 (%)	83.00	75.00	68.04	34.82	75.89	71.43	79.46	78.57	68.04	80.36
Ionosphere	d	29	31	8	8	16	11	28	34	33	9
	SVM (%)	87.75	87.18	74.93	75.49	86.32	81.48	86.89	84.62	85.19	87.75
	C4.5 (%)	92.60	95.20	88.30	92.02	95.20	91.50	97.40	95.44	97.44	98.00
Segmentation	d	18	18	14	9	13	12	15	17	15	10
	SVM (%)	90.43	90.43	91.57	73.91	92.19	90.52	91.05	91.38	91.43	95.24
	C4.5 (%)	90.38	90.24	90.00	77.91	89.38	89.24	90.00	86.05	85.76	90.00
Lung	d	2	3	7	6	10	7	2	4	4	13
	SVM (%)	59.06	57.72	63.08	71.14	74.49	74.49	80.54	81.21	82.55	87.92
	C4.5 (%)	74.50	74.50	71.82	80.53	77.85	71.82	70.50	77.85	81.21	85.91
DLBCLNIH	d	94	81	17	12	22	29	83	72	63	19
	SVM (%)	55.00	60.00	60.00	56.25	61.25	58.75	65.00	70.00	70.00	70.00
	C4.5 (%)	48.75	57.50	61.25	60.00	57.50	58.75	62.50	61.25	68.75	68.75
Prostate	d	21	33	4	7	6	9	23	18	15	5
	SVM (%)	82.35	88.24	76.47	79.41	82.35	85.29	88.23	91.18	91.18	94.12
	C4.5 (%)	67.65	79.41	67.65	76.47	79.41	67.65	85.29	88.24	94.12	94.12
Multiple Features	d	61	44	11	10	15	12	33	57	48	18
	SVM (%)	88.21	83.90	79.75	78.50	83.90	81.30	84.95	81.80	85.75	90.68
	C4.5 (%)	88.21	88.60	76.70	76.70	86.45	80.35	80.35	84.95	88.90	88.21

TABLE 7
Comparative Execution Time (in milli second) of Different Feature Selection Methods

Data Sets	InfoGain	mRMR	RSQR	RSMRMS	FRQR	FRmRMR	RELIEF	SIMBA	Proposed/PAA
Breast	1833	24542	235651	13190	790851	25714	387	613	1344/829
Satimage	227	1247	38987	981	42178	1341	117	201	462/236
Isolet	14139	133508	70488913	108973	71412733	139733	8849	10911	107101/55837
Leukemia	11958	110519	117387083	53796	526377189	116548	6329	8273	14654/6579
Ionosphere	59	88	2107	47	66487	95	10	11	11/6
Segmentation	26	46	167998	21	170276	35	7	7	11/5
Lung	636	1090	3512	971	4061	1107	511	523	157/82
DLBCLNIH	21061	35814	372948	40761	436061	105874	10754	11037	651/289
Prostate	14678	22997	6303	4173	1312773	16984	13728	17894	272/153
Multiple Features	8917	9043	574289	96216	649239	11065	9281	10294	461/279

evance in estimating features (RELIEF) [27] and iterative search margin based algorithm (SIMBA) [36].

From the results reported in Table 6, it is seen that the proposed method selects a set of features having highest accuracy of both SVM and C4.5 in most of the cases. Also, the proposed method can potentially yield significantly better results than existing algorithms. The method due to Chen and Wasikowski [37] for binary class data sets achieves 63.16%, 81.48%, 80.54%, 68.75% and 85.29% accuracy using the SVM, and 52.63%, 92.02%, 80.53%, 61.25% and 88.24% accuracy using the C4.5 on Breast Cancer, Ionosphere, Lung, DLBCLNIH, Prostate Cancer data, respectively. Fig. 3 presents the variation of classification error of the SVM over different number of selected features on Satimage and Segmentation data. All the results reported in Fig. 3 establish that the proposed method significantly outperforms others, especially small number of features regime. The better performance of hypercuboid equivalence partition matrix based proposed method is achieved due to the fact that it provides an efficient way to compute degree of dependency of class labels on feature set in approximation spaces. In effect, a reduced set of relevant and significant features is being obtained using the proposed method.

Moreover, Table 7 reports the execution time of different algorithms. The significantly lesser time of the proposed algorithm is achieved due to its low computational complexity. The execution time of the proposed algorithm is reduced significantly when it is implemented using (28) based on the concept of positive approximation accelerator (PAA) [30], and the difference is more visible for large data sets, both in size and dimension.

7 Conclusion

The contribution of the paper is three fold, namely, the development of a feature selection algorithm, integrating

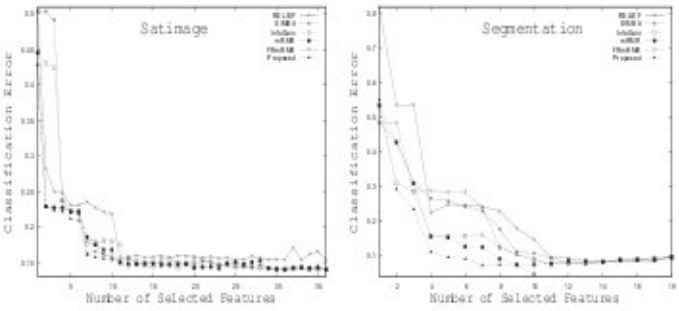

Fig. 3. Classification error of the SVM obtained using various methods for different number of selected features

the merits of rough sets and hypercuboid equivalence partition matrix; defining new quantitative indices based on rough hypercuboid approach in order to describe the inexactness of approximate classification; and demonstrating the effectiveness of the proposed algorithm, along with a comparison with other algorithms, on several real life data sets.

The concept of hypercuboid equivalence partition matrix is found to be successful in selecting relevant and significant features of real valued data sets. This formulation is geared towards maximizing the utility of rough sets and hypercuboid approach with respect to knowledge discovery tasks. The methodology of integrating rough sets and hypercuboid approach can also be applied to other feature selection problems, and the proposed indices may be used in a suitable combination to act as the objective function of an evolutionary algorithm for feature selection.

REFERENCES

[1] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification and Scene Analysis*. New York: John Wiley & Sons, Inc., 1999.

[2] P. A. Devijver and J. Kittler, *Pattern Recognition: A Statistical Approach*. Englewood Cliffs: Prentice Hall, 1982.

[3] Z. Pawlak, *Rough Sets: Theoretical Aspects of Reasoning About Data*. Dordrecht, The Netherlands: Kluwer, 1991.

[4] L. Polkowski, *Rough Sets*. Heidelberg: Physica-Verlag, 2002.

[5] A. Skowron and C. Rauszer, "The Discernibility Matrices and Functions in Information Systems," in *Intelligent Decision Support*, R. Slowinski, Ed. Dordrecht: Kluwer Academic Publishers, 1992, pp. 331–362.

[6] A. Skowron, "Extracting Laws from Decision Tables: A Rough Set Approach," *Computational Intelligence*, vol. 11, pp. 371–388, 1995.

[7] A. Chouchoulas and Q. Shen, "Rough Set-Aided Keyword Reduction for Text Categorisation," *Applied Artificial Intelligence*, vol. 15, no. 9, pp. 843–873, 2001.

[8] A. Skowron and S. K. Pal, "Rough Sets, Pattern Recognition, and Data Mining," *Pattern Recognition Letters*, vol. 24, no. 6, pp. 829–933, 2003.

[9] R. W. Swiniarski and A. Skowron, "Rough Set Methods in Feature Selection and Recognition," *Pattern Recognition Letters*, vol. 24, pp. 833–849, 2003.

[10] N. Parthalain, Q. Shen, and R. Jensen, "A Distance Measure Approach to Exploring the Rough Set Boundary Region for Attribute Reduction," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 3, pp. 305–317, 2010.

[11] S. H. Nguyen and A. Skowron, "Quantization of Real Value Attributes - Rough Set and Boolean Reasoning Approach," in *Proceedings of the Second Joint Annual Conference on Information Sciences, North Carolina*, 1995, pp. 34–37.

[12] D. Dubois and H. Prade, "Rough Fuzzy Sets and Fuzzy Rough Sets," *International Journal of General Systems*, vol. 17, pp. 191–209, 1990.

[13] S. K. Pal and A. Skowron, Eds., *Rough-Fuzzy Hybridization: A New Trend in Decision Making*. Singapore: Springer-Verlag, 1999.

[14] D. S. Yeung, D. Chen, E. C. C. Tsang, J. W. T. Lee, and W. Xizhao, "On the Generalization of Fuzzy Rough Sets," *IEEE Transactions on Fuzzy Systems*, vol. 13, no. 3, pp. 343–361, 2005.

[15] X. Liu, W. Pedrycz, T. Chai, and M. Song, "The Development of Fuzzy Rough Sets with the Use of Structures and Algebras of Axiomatic Fuzzy Sets," *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 3, pp. 443–462, 2009.

[16] P. Maji and S. K. Pal, *Rough-Fuzzy Pattern Recognition: Applications in Bioinformatics and Medical Imaging*. New Jersey, Hoboken: John Wiley & Sons, Inc., 2012.

[17] R. Jensen and Q. Shen, "Semantics-Preserving Dimensionality Reduction: Rough and Fuzzy-Rough-Based Approach," *IEEE Transactions on Knowledge and Data Engineering*, vol. 16, no. 12, pp. 1457–1471, 2004.

[18] Q. Hu, D. Yu, Z. Xie, and J. Liu, "Fuzzy Probabilistic Approximation Spaces and Their Information Measures," *IEEE Transactions on Fuzzy Systems*, vol. 14, no. 2, pp. 191–201, 2007.

[19] R. Jensen and Q. Shen, "Fuzzy-Rough Sets Assisted Attribute Selection," *IEEE Transactions on Fuzzy Systems*, vol. 15, pp. 73–89, 2007.

[20] P. Maji and S. K. Pal, "Feature Selection Using f-Information Measures in Fuzzy Approximation Spaces," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 6, pp. 854–867, 2010.

[21] E. C. C. Tsang, D. Chen, D. S. Yeung, X.-Z. Wang, and J. Lee, "Attributes Reduction Using Fuzzy Rough Sets," *IEEE Transactions on Fuzzy Systems*, vol. 16, no. 5, pp. 1130–1141, 2008.

[22] R. Jensen and Q. Shen, "New Approaches to Fuzzy-Rough Feature Selection," *IEEE Transactions on Fuzzy Systems*, vol. 17, no. 4, pp. 824–838, 2009.

[23] Q. Hu, D. Yu, J. Liu, and C. Wu, "Neighborhood Rough Set Based Heterogeneous Feature Subset Selection," *Information Sciences*, vol. 178, pp. 3577–3594, 2008.

[24] J.-M. Wei, S.-Q. Wang, and X.-J. Yuan, "Ensemble Rough Hypercuboid Approach for Classifying Cancers," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 3, pp. 381–391, 2010.

[25] P. Maji and S. Paul, "Rough Set Based Maximum Relevance-Maximum Significance Criterion and Gene Selection from Microarray Data," *International Journal of Approximate Reasoning*, vol. 52, no. 3, pp. 408–426, 2011.

[26] J. R. Quinlan, "Induction of Decision Trees," *Machine Learning*, vol. 1, pp. 81–106, 1986.

[27] K. Kira and L. A. Rendell, "The Feature Selection Problem: Traditional Methods and A New Algorithm," in *Proceedings of the 10th National Conference on Artificial Intelligence*. MIT Press, 1992, pp. 129–134.

[28] I. Guyon and A. Elisseeff, "An Introduction to Variable and Feature Selection," *Journal of Machine Learning Research*, vol. 3, pp. 1157–1182, 2003.

[29] H. Peng, F. Long, and C. Ding, "Feature Selection Based on Mutual Information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 8, pp. 1226–1238, 2005.

[30] Y. Qian, J. Liang, W. Pedrycz, and C. Dang, "Positive Approximation: An Accelerator for Attribute Reduction in Rough Set Theory," *Artificial Intelligence*, vol. 174, pp. 597–618, 2010.

[31] A. Frank and A. Asuncion, "UCI Machine Learning Repository," 2010. [Online]. Available: http://archive.ics.uci.edu/ml

[32] "Kent Ridge Bio-medical Data Set Repository." [Online]. Available: http://sdmc.lit.org.sg/GEDatasets/Datasets.html

[33] M. Dash and H. Liu, "Unsupervised Feature Selection," in *Proceedings of the Pacific Asia Conference on Knowledge Discovery and Data Mining*, 2000, pp. 110–121.

[34] V. Vapnik, *The Nature of Statistical Learning Theory*. New York: Springer-Verlag, 1995.

[35] J. R. Quinlan, *C4.5: Programs for Machine Learning*. CA: Morgan Kaufmann, 1993.

[36] R. Gilad-Bachrach, A. Navot, and N. Tishby, "Margin Based Feature Selection: Theory and Algorithms," in *Proceedings of the 21st International Conference on Machine Learning*, 2004.

[37] X.-W. Chen and M. Wasikowski, "FAST: A ROC-Based Feature Selection Metric for Small Samples and Imbalanced Data Classification Problems," in *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2008.