Molecular BioSystems

PAPER

View Article Online
View Journal | View Issue

Cite this: *Mol. BioSyst.*, 2015, 11, 2068

Rough hypercuboid based supervised clustering of miRNAs

Sushmita Paul* and Julio Vera

Received 26th March 2015,
Accepted 14th May 2015

DOI: 10.1039/c5mb00213c

www.rsc.org/molecularbiosystems

The microRNAs are small, endogenous non-coding RNAs found in plants, animals, and some viruses, which function in RNA silencing and post-transcriptional regulation of gene expression. It is suggested by various genome-wide studies that a substantial fraction of miRNA genes is likely to form clusters. The coherent expression of the miRNA clusters can then be used to classify samples according to the clinical outcome. In this regard, a new clustering algorithm, termed as rough hypercuboid based supervised attribute clustering (RH-SAC), is proposed to find such groups of miRNAs. The proposed algorithm is based on the theory of rough set, which directly incorporates the information of sample categories into the miRNA clustering process, generating a supervised clustering algorithm for miRNAs. The effectiveness of the new approach is demonstrated on several publicly available miRNA expression data sets using support vector machine. The so-called *B*.632+ bootstrap error estimate is used to minimize the variability and biasedness of the derived results. The association of the miRNA clusters to various biological pathways is also shown by doing pathway enrichment analysis.

1 Introduction

MicroRNA/miRNAs are a class of short approximately 22-nucleotide non-coding RNAs processed from hairpin precursors of ~70 nt (pre-miRNA), extracted, in turn, from primary transcripts (pri-miRNA) found in many plants and animals. Their roles have been studied in many crucial biological processes, including development, differentiation, apoptosis and cell proliferation,[1–4] as well as numerous human diseases, such as chronic lymphocytic leukemia, fragile X syndrome, and various types of cancers.[5–8] The binding of miRNAs to the 3′ untranslated region of the mRNA leads to the down regulation of that mRNA expression.

It has been shown by ref. 9 that the miRNAs on a genome tend to present in a cluster. However, the evolutionary and biological function implications of clustered miRNAs are still elusive. In ref. 10, a clustering algorithm is described to identify miRNA clusters. Chan *et al.*[11] used a data mining approach to discover miRNA cluster patterns. According to Wang *et al.*,[12] the set of miRNAs that are closely distributed in genome is termed as the miRNA cluster. Large scale surveys[13,14] have established the fact that miRNAs have tendency to present in clusters. Based on these studies, it is confirmed that miRNA clusters are widely distributed in animal genomes.[15,16] The miRNA clusters adapt special regulatory functions in biological processes due to the conservation of miRNA clusters across species.[9] It has been reported that at a very conservative maximum inter-miRNA distance of 1 kb, over 30% of all miRNAs are organized into clusters.[9] Expression analyses showed strong positive correlations among the closely located miRNAs, indicating that they may be controlled by common regulatory element(s). In fact, experimental evidence demonstrated that clustered miRNA loci form an operon-like gene structure and that they are transcribed from common promoter. Hence, it is assumed that the genomic coordination of clustered miRNA genes, which further leads to their coordinated transcription, will consequently result in a functional coordination.[12] Existence of co-expressed miRNAs is also demonstrated using expression profiling analysis in ref. 17. Several miRNA clusters have been experimentally shown by RT-PCR or Northern blotting.[18,19] These findings suggest that members of a miRNA cluster, which are at a close proximity on a chromosome, are highly likely to be processed as co-transcribed units. Expression data of miRNAs can be used to detect clusters of miRNAs as it is suggested that co-expressed miRNAs are co-transcribed, so they should have similar expression pattern.

Few methods have been employed to identify differentially expressed miRNAs. Most of the analysis are based on statistical tests.[20–29] In few works, significance analysis of microarrays is used to select miRNAs from its expression data.[30–35] Along with expression data, sequence data of miRNAs are used to identify potential miRNAs. In ref. 36, Xue *et al.* developed a method, termed as Triplet-SVM, that uses sequence data to classify real pre-miRNAs and pseudo pre-miRNAs using support vector machine. Computational methods like miRNA-dis,[37] iMcRNA,[38]

Laboratory of Systems Tumor Immunology, Department of Dermatology, University of Erlangen-Nürnberg, Hartmannstr. 14, 91052 Erlangen, Germany. E-mail: sushmita.paul@uk-erlangen.de, julio.vera-gonzalez@uk-erlangen.de

and iMiRNA-PseDPC[39] are also developed to identify microRNA precursor using sequence data.

Several unsupervised clustering techniques have been used to cluster a miRNA expression data. Several authors used hierarchical clustering algorithms,[20,28,40] self organizing maps,[41] and rough-fuzzy clustering algorithms[10] to group miRNAs having similar function. Other clustering techniques such as *k*-means algorithm,[42] graph theoretical approaches,[43–46] model based clustering,[47–50] rough set based clustering algorithms,[10,51] and density based approach,[52] which have been widely applied to find co-expressed gene clusters, can also be used to group co-expressed miRNAs from microarray data. However, the groups of miRNAs discovered by all these unsupervised clustering algorithms are not potential enough to do tissue classification,[53] as the miRNAs are grouped based on their similarity without incorporating the class label information.

In this regard, several supervised clustering algorithms are proposed to cluster gene expression data.[53–56] In ref. 53, genes are clustered by incorporating the knowledge of tissue. On the other hand, hierarchical clustering is employed on the gene expression data and the average of resultant clustering solutions are further used to do sample classification. Only in the later part, information of the class label is incorporated.[56] In ref. 55, a fuzzy-rough supervised gene clustering algorithm is described, which uses fuzzy equivalence classes to compute relevance of the clusters, that makes the algorithm sensitive to the fuzzy parameter. However, none of the works has addressed the problem of supervised clustering of miRNAs.

Also, one of the main problems in expression data analysis is uncertainty. Some of the sources of this uncertainty include imprecision in computations and vagueness in class definition. In this background, the rough set[57] provides a mathematical framework to capture uncertainties associated with human cognition process.[58,59] In ref. 60–62, rough sets have been successfully used to identify differentially expressed genes from gene expression data. Importance of rough sets is also shown in clustering analysis. Rough sets are used to design clustering algorithms,[51,63] to identify groups of co-expressed genes from gene microarray data sets.

In this regard, this paper presents a new rough hypercuboid based supervised clustering algorithm. It is developed by integrating the concepts of rough hypercuboid equivalence partition matrix[62,64] and supervised attribute clustering algorithm.[55] It finds coregulated clusters of miRNAs whose collective expression is strongly associated with the sample categories. Using the concept of rough hypercuboid equivalence partition matrix, the degree of dependency is calculated for miRNAs, which is used to compute both relevance and significance of the miRNAs. Hence, the only information required in the proposed method is in the form of equivalence classes for each miRNA, which can be automatically derived from the data set. A new measure is introduced for calculating similarity between two miRNAs. Based upon the similarity values, the miRNAs are grouped into cluster. The proposed supervised clustering algorithm divides the miRNA expression data into distinct clusters. In each cluster, the first selected miRNA has high relevance value with respect to the class label and it is the representative of the cluster. The representative is modified in such a way that the averaged expression value has high relevance value with the class label. Finally, the proposed method generates a set of clusters, whose coherent average expression levels allow perfect discrimination of tissue types. The concept of *B*.632+ error rate[65] is used to minimize the variability and biasedness of the derived results. The support vector machine is used to compute the *B*.632+ error rate as well as several other types of error rates as it maximizes the margin between data samples in different classes. The effectiveness of the proposed approach, along with a comparison with other related approaches, is demonstrated on several miRNA expression data sets.

The structure of the rest of this paper is as follows: Section 2 briefly introduces rough set theory. Section 3 describes the rough hypercuboid equivalence partition matrix. The supervised similarity measure is also discussed in this section, along with the proposed rough hypercuboid based supervised miRNA clustering algorithm. Finally, *B*.632+ error rate, support vector machine, and important steps of the proposed method are also described in this section. A few case studies and a comparison with existing algorithms are presented in Section 4. Concluding remarks are given in Section 5.

2 Rough sets

The proposed supervised miRNA clustering algorithm is developed by using the concept of rough sets. This section describes the basic concepts of rough sets. Let $\langle \mathbb{U},\mathbb{M} \rangle$ be an approximation space or an information system, where $\mathbb{U} = \{s_1,\ldots,s_i,\ldots,s_n\}$ be a non-empty set, the universe of discourse and $\mathbb{M}$ is a family of attributes or miRNAs, also called knowledge in the universe. V is the value domain of $\mathbb{M}$ and f is an information function f: $\mathbb{U} \times \mathbb{M} \rightarrow V$.[57] Any subset $\mathbb{P}$ of knowledge $\mathbb{M}$ defines an equivalence or indiscernibility relation IND($\mathbb{P}$) on $\mathbb{U}$

$$\mathrm{IND}(\mathbb{P}) = \{(s_i,s_j) \in \mathbb{U} \times \mathbb{U} | \forall a \in \mathbb{P}, f_a(s_i) = f_a(s_j)\}.$$

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

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

where $[s_i]_{\mathbb{P}}$ is the equivalence class containing s_i. The elements in $[s_i]_{\mathbb{P}}$ 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 $\varnothing$ are the elementary sets in the approximation space $\langle \mathbb{U},\mathbb{M} \rangle$.

Given an arbitrary set $X \subseteq \mathbb{U}$, in general, it may not be possible to describe X precisely in $\langle \mathbb{U},\mathbb{M} \rangle$. One may characterize X by a pair of lower and upper approximations defined as follows:[57]

$$\underline{\mathbb{P}}(X) = \bigcup \{[s_i]_{\mathbb{P}} | [s_i]_{\mathbb{P}} \subseteq X\} \text{ and}$$

$$\overline{\mathbb{P}}(X) = \bigcup \{[s_i]_{\mathbb{P}} | [s_i]_{\mathbb{P}} \cap X \neq \varnothing\}.$$

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 $\langle \underline{\mathbb{P}}(X), \overline{\mathbb{P}}(X) \rangle$ is the representation of an ordinary set X in the approximation space $\langle \mathbb{U}, \mathbb{M} \rangle$ or simply called the rough sets 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 $\mathrm{POS}_{\mathbb{P}}(X)$. A set X is said to be definable or exact in $\langle \mathbb{U}, \mathbb{M} \rangle$ iff $\underline{\mathbb{P}}(X) = \overline{\mathbb{P}}(X)$. Otherwise X is indefinable and termed as a rough set.[57]

Definition 1. An information system $\langle \mathbb{U}, \mathbb{M} \rangle$ is called a decision table if the attribute set $\mathbb{M} = \mathbb{C} \cup \mathbb{D}$, where $\mathbb{C}$ and $\mathbb{D}$ represent the condition and decision attribute sets, respectively. The dependency between $\mathbb{C}$ and $\mathbb{D}$ can be defined as[57]

$$\gamma_{\mathbb{C}}(\mathbb{D}) = \frac{|\mathrm{POS}_{\mathbb{C}}(\mathbb{D})|}{|\mathbb{U}|} \tag{2}$$

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

Definition 2. The change in dependency when an attribute/miRNA is removed from the set of condition attributes/miRNAs, is a measure of the significance of the attribute or miRNA. To what extent an attribute is contributing to calculate the dependency on decision attribute can be calculated by the significance of that attribute. Given $\mathbb{C}$, $\mathbb{D}$ and an attribute $\mathscr{M} \in \mathbb{C}$, the significance of the attribute $\mathscr{M}$ is defined as:[57]

$$\sigma_{\mathbb{C}}(\mathbb{D}, \mathscr{M}) = \gamma_{\mathbb{C}}(\mathbb{D}) - \gamma_{\mathbb{C}-\{\mathscr{M}\}}(\mathbb{D}). \tag{3}$$

The nearer the value of attribute/miRNA $\mathscr{M}$ is to 1, the more it is significant.

3 Proposed rough hypercuboid based supervised attribute clustering

In this paper, a new clustering algorithm is developed based on rough hypercuboid equivalence partition matrix.[62,64] The concept of rough hypercuboid was initially introduced in ref. 64 and 66 and it was successfully used in ref. 62. A new rough hypercuboid based similarity measure is proposed, as every clustering algorithm need a distance or similarity measure to group objects. The relevance of a cluster is also calculated using rough hypercuboid based dependency measure. Prior to describe the proposed supervised attribute clustering algorithm, next the concept of rough hypercuboid equivalence partition matrix[62,64] is described.

3.1 Rough hypercuboid equivalence partition matrix

Let $\mathbb{U} = \{s_1, \ldots, s_i, \ldots, s_n\}$ be the set of n objects or samples and $\mathbb{C} = \{\mathscr{M}_1, \ldots, \mathscr{M}_i, \ldots, \mathscr{M}_j, \ldots, \mathscr{M}_m\}$ denotes the set of m attributes or miRNAs of a given microarray data set $\mathscr{T} = \{w_{ij} | i = 1, \ldots, m, j = 1, \ldots, n\}$, where $w_{ij} \in \Re$ is the measured expression value of the miRNA $\mathscr{M}_i$ in the sample s_j. Let $\mathbb{D}$ be the set of class labels or sample categories of n samples.

If $\mathbb{U}/\mathbb{D} = \{\beta_1, \ldots, \beta_i, \ldots, \beta_c\}$ 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 or miRNA $\mathscr{M}_k \in \mathbb{C}$. If $\mathbb{U}/\mathscr{M}_k = \{\mu_1, \ldots, \mu_i, \ldots, \mu_c\}$ denotes c equivalence classes or information granules of $\mathbb{U}$ induced by the condition attribute or miRNA $\mathscr{M}_k$ and n is the number of objects in $\mathbb{U}$, then c-partitions of $\mathbb{U}$ are the sets of (cn) values $\{\mathrm{h}_{ij}(\mathscr{M}_k)\}$ that can be conveniently arrayed as a $(c \times n)$ matrix $\mathbb{H}(\mathscr{M}_k) = [\mathrm{h}_{ij}(\mathscr{M}_k)]$.[62,64] The matrix $\mathbb{H}(\mathscr{M}_k)$ is denoted by

$$\mathbb{H}(\mathscr{M}_k) = \begin{pmatrix} \mathrm{h}_{11}(\mathscr{M}_k) & \mathrm{h}_{12}(\mathscr{M}_k) & \cdots & \mathrm{h}_{1n}(\mathscr{M}_k) \\ \mathrm{h}_{21}(\mathscr{M}_k) & \mathrm{h}_{22}(\mathscr{M}_k) & \cdots & \mathrm{h}_{2n}(\mathscr{M}_k) \\ \cdots & \cdots & \cdots & \cdots \\ \mathrm{h}_{c1}(\mathscr{M}_k) & \mathrm{h}_{c2}(\mathscr{M}_k) & \cdots & \mathrm{h}_{cn}(\mathscr{M}_k) \end{pmatrix} \tag{4}$$

$$\text{where } \mathrm{h}_{ij}(\mathscr{M}_k) = \begin{cases} 1 & \text{if } \mathrm{L}_i \le x_j(\mathscr{M}_k) \le \mathrm{U}_i \\ 0 & \text{otherwise} \end{cases}. \tag{5}$$

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 or miRNA $\mathscr{M}_k$ 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 s_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.

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.

On employing a condition attribute or miRNA $\mathscr{M}_k$, a $c \times n$ matrix $\mathbb{H}(\mathscr{M}_k)$, termed as hypercuboid equivalence partition matrix, is generated. It represents the c-hypercuboid equivalence partitions of the universe generated by an equivalence relation. Each row of the matrix $\mathbb{H}(\mathscr{M}_k)$ is a hypercuboid equivalence partition or class. Here $\mathrm{h}_{ij}(\mathscr{M}_k) \in \{0,1\}$ represents the membership of object s_j in the ith equivalence partition or class β_i satisfying following two conditions:

$$1 \le \sum_{j=1}^{n} \mathrm{h}_{ij}(\mathscr{M}_k) \le n, \quad \forall i; \tag{6}$$

$$1 \leq \sum_{i=1}^{c} \mathrm{h}_{ij}(\mathscr{M}_k) \leq c, \quad \forall j. \tag{7}$$

The above axioms should hold for every equivalence partition, which correspond to the requirement that an equivalence class is non-empty. However, in real data analysis, uncertainty arises due to overlapping class boundaries. Hence, such a granulation process does not necessarily result in a compatible granulation in the sense that every two class hypercuboids or intervals 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. Hence, the degree of dependency of decision attribute on a condition attribute set 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[62,64]

$$\mathbb{V}(\mathscr{M}_k) = [v_1(\mathscr{M}_k),\ldots,v_j(\mathscr{M}_k),\ldots,v_n(\mathscr{M}_k)] \tag{8}$$

$$\text{where } v_j(\mathscr{M}_k) = \min\left\{1, \sum_{i=1}^{c} \mathrm{h}_{ij}(\mathscr{M}_k) - 1\right\}. \tag{9}$$

As already mentioned that if an object s_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 $v_j(\mathscr{M}_k) = 0$. On the other hand, if the object s_j belongs to the boundary region of more than one classes, then it should be encompassed by the implicit hypercuboid and $v_j(\mathscr{M}_k) = 1$. Hence, the hypercuboid equivalence partition matrix and corresponding confusion vector of the condition attribute $\mathscr{M}_k$ 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 $\mathscr{M}_k$ by constructing the M-lower and M-upper approximations of β_i:[62,64]

$$\underline{M}(\beta_i) = \{s_j | \mathrm{h}_{ij}(\mathscr{M}_k) = 1 \text{ and } v_j(\mathscr{M}_k) = 0\}; \tag{10}$$

$$\bar{M}(\beta_i) = \{s_j | \mathrm{h}_{ij}(\mathscr{M}_k) = 1\}; \tag{11}$$

where equivalence relation M is induced from attribute $\mathscr{M}_k$. The boundary region of β_i is then defined as[62,64]

$$\mathrm{BN}_M(\beta_i) = \{s_j | \mathrm{h}_{ij}(\mathscr{M}_k) = 1 \text{ and } v_j(\mathscr{M}_k) = 1\}. \tag{12}$$

Dependency. Combining (4), (8), and (10), the dependency between condition attribute $\mathscr{M}_k$ and decision attribute $\mathbb{D}$ can be defined as follows:[62,64]

$$\gamma_{\mathcal{M}_k}(\mathbb{D}) = \frac{1}{n}\sum_{i=1}^{c}\sum_{j=1}^{n} h_{ij}(\mathscr{M}_k) \cap \left[1 - v_j(\mathscr{M}_k)\right], \tag{13}$$

$$\text{that is, } \gamma_{\mathscr{M}_k}(\mathbb{D}) = 1 - \frac{1}{n}\sum_{j=1}^{n} v_j(\mathscr{M}_k), \tag{14}$$

where $0 \leq \gamma_{\mathscr{M}_k}(\mathbb{D}) \leq 1$. If $\gamma_{\mathscr{M}_k}(\mathbb{D}) = 1$, $\mathbb{D}$ depends totally on $\mathscr{M}_k$, if $0 < \gamma_{\mathscr{M}_k}(\mathbb{D}) < 1$, $\mathbb{D}$ depends partially on $\mathscr{M}_k$, and if $\gamma_{\mathscr{M}_k}(\mathbb{D}) = 0$, then $\mathbb{D}$ does not depend on $\mathscr{M}_k$. The $\gamma_{\mathscr{M}_k}(\mathbb{D})$ is also termed as the relevance of attribute $\mathscr{M}_k$ with respect to class $\mathbb{D}$.

Significance. Given two condition attributes or miRNAs $\mathscr{M}_k$ and $\mathscr{M}_l$, the $c \times n$ hypercuboid equivalence partition matrix corresponding to the set $\mathbb{M} = \{\mathscr{M}_k,\mathscr{M}_l\}$ can be calculated from two $c \times n$ hypercuboid equivalence partition matrices $\mathbb{H}(\mathscr{M}_k)$ and $\mathbb{H}(\mathscr{M}_l)$ as follows:

$$\mathbb{H}(\{\mathscr{M}_k,\mathscr{M}_l\}) = \mathbb{H}(\mathscr{M}_k) \cap \mathbb{H}(\mathscr{M}_l); \tag{15}$$

$$\text{where} \quad \mathrm{h}_{ij}(\{\mathscr{M}_k,\mathscr{M}_l\}) = \mathrm{h}_{ij}(\mathscr{M}_k) \cap \mathrm{h}_{ij}(\mathscr{M}_l). \tag{16}$$

The significance of the attribute $\mathscr{M}_k$ with respect to the condition attribute set $\{\mathscr{M}_k,\mathscr{M}_l\}$ is given by[62,64]

$$\sigma_{\mathbb{M}}(\mathbb{D},\mathscr{M}_k) = \frac{1}{n}\sum_{j=1}^{n}\left[v_j(\mathbb{M} - \{\mathscr{M}_k\}) - v_j(\mathbb{M})\right]; \tag{17}$$

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

As already mentioned, a distance or similarity measure is required for grouping two objects in any cluster analysis technique. In this regard, a new rough hypercuboid based similarity measure is developed in this work. The similarity measure uses the information of class labels. Hence, it is a supervised similarity measure. The following subsection describes the proposed similarity measure in detail.

3.2 Rough hypercuboid based supervised similarity measure

The simple concept of rough hypercuboid based dependency and significance is used to calculate dissimilarity between two miRNAs and then the non-linear transformation of the dissimilarity is used to calculate similarity between two miRNAs. This subsection presents the proposed rough hypercuboid based supervised similarity measure.

In real data analysis, the functional relationship between a biomarker and the clinical outcome can be established by computing the relevance and redundancy of biomarkers with respect to the clinical outcome. Intuitively, a set of attributes $\mathbb{Q}$ depends totally on a set of attributes $\mathbb{P}$, if all attribute values from $\mathbb{Q}$ are uniquely determined by values of attributes from $\mathbb{P}$. If there exists a functional dependency between values of $\mathbb{Q}$ and $\mathbb{P}$, then $\mathbb{Q}$ depends totally on $\mathbb{P}$.

Let $\mathbb{C} = \{\mathscr{M}_1,\ldots,\mathscr{M}_i,\ldots,\mathscr{M}_j,\ldots,\mathscr{M}_m\}$ denotes the set of m condition attributes or miRNAs of a given data set. Define $\mathrm{R}_{\mathscr{M}_i}(\mathbb{D})$ as the relevance of the condition attribute $\mathscr{M}_i$ with respect to the class label or decision attribute $\mathbb{D}$. The dependency function of rough hypercuboid can be used to calculate the relevance of condition attributes or miRNAs. Hence, the relevance $\mathrm{R}_{\mathscr{M}_i}(\mathbb{D})$ of the condition attribute $\mathscr{M}_i$ with respect to the decision attribute $\mathbb{D}$ using rough hypercuboid can be calculated as follows:

$$\mathrm{R}_{\mathscr{M}_i}(\mathbb{D}) = \gamma_{\mathscr{M}_i}(\mathbb{D}) \tag{18}$$

where $\gamma_{\mathscr{M}_i}(\mathbb{D})$ represents the degree of dependency between condition attribute or miRNA $\mathscr{M}_i$ and decision attribute or class label $\mathbb{D}$ that is given by (14).

At first, the dissimilarity between two miRNAs $\mathscr{M}_i$ and $\mathscr{M}_j$ is calculated using rough hypercuboid based approach. Then the non-linear transformation of the dissimilarity is done for getting the similarity between these two miRNAs. The non-linear transformation is done to detect nonlinear interdependencies between the underlying two miRNAs. The rough hypercuboid based significance (17) is used to compute similarity between two miRNAs and it is defined next.

Definition 3. The rough hypercuboid based similarity measure between two attributes or miRNAs $\mathscr{M}_i$ and $\mathscr{M}_j$ is defined as follows:

$$\psi(\mathscr{M}_i, \mathscr{M}_j) = \frac{1}{\sqrt{\kappa^2 + 1}}, \tag{19}$$

$$\text{where } \kappa = \left\{ \frac{\sigma_{\{\mathscr{M}_i,\mathscr{M}_j\}}(\mathbb{D}, \mathscr{M}_j) + \sigma_{\{\mathscr{M}_i,\mathscr{M}_j\}}(\mathbb{D}, \mathscr{M}_i)}{2} \right\} \tag{20}$$

$$\text{that is, } \kappa = \mathrm{R}_{\{\mathscr{M}_i,\mathscr{M}_j\}}(\mathbb{D}) - \left\{ \frac{\mathrm{R}_{\mathscr{M}_i}(\mathbb{D}) + \mathrm{R}_{\mathscr{M}_j}(\mathbb{D})}{2} \right\}. \tag{21}$$

Hence, the supervised similarity measure $\psi(\mathscr{M}_i,\mathscr{M}_j)$ directly takes into account the information of sample categories or class labels $\mathbb{D}$ while computing the similarity between two attributes or miRNAs $\mathscr{M}_i$ and $\mathscr{M}_j$. If attributes $\mathscr{M}_i$ and $\mathscr{M}_j$ are completely correlated with respect to class labels $\mathbb{D}$, then $\kappa = 0$ and so $\psi(\mathscr{M}_i,\mathscr{M}_j)$ is 1. If $\mathscr{M}_i$ and $\mathscr{M}_j$ are totally uncorrelated, $\psi(\mathscr{M}_i, \mathscr{M}_j) = \frac{1}{\sqrt{2}}$. Hence, $\psi(\mathscr{M}_i,\mathscr{M}_j)$ can be used as a measure of supervised similarity between two attributes $\mathscr{M}_i$ and $\mathscr{M}_j$. The following properties can be stated about the measure:

(1) $\frac{1}{\sqrt{2}} \leq \psi(\mathscr{M}_i, \mathscr{M}_j) \leq 1$.

(2) $\psi(\mathscr{M}_i,\mathscr{M}_j) = 1$ if and only if $\mathscr{M}_i$ and $\mathscr{M}_j$ are completely correlated.

(3) $\psi(\mathscr{M}_i, \mathscr{M}_j) = \frac{1}{\sqrt{2}}$ if and only if $\mathscr{M}_i$ and $\mathscr{M}_j$ are totally uncorrelated.

(4) $\psi(\mathscr{M}_i,\mathscr{M}_j) = \psi(\mathscr{M}_j,\mathscr{M}_i)$ (symmetric).

Therefore, the rough hypercuboid based similarity measure $\psi(\mathscr{M}_i,\mathscr{M}_j)$ between two attributes or miRNAs $\mathscr{M}_i$ and $\mathscr{M}_j$ can be used to compute the redundancy among the attributes taking into account the information of class label while computing the similarity between two attributes or miRNAs.

3.3 Supervised miRNA clustering algorithm

In this work, the proposed rough hypercuboid based similarity measure is incorporated into the fuzzy-rough supervised attribute clustering algorithm.[55] In the proposed method, a new rough hypercuboid based similarity measure is developed to calculate similarity between two miRNAs. Whereas, in ref. 55, a fuzzy-rough supervised similarity measure is proposed. However, the fuzzy-rough supervised similarity measure is sensitive to the fuzzy parameter that is used to calculate the similarity between two objects. In supervised miRNA clustering algorithm, initially the most relevant attribute/miRNA is selected. Then, the cluster is grown incrementally by adding one attribute/miRNA after the other. Once the growth of the clustering algorithm gets stabilized, that means the cluster has more identical miRNAs and the averaged expression values of the clustered miRNAs are differentially expressed, then the supervised miRNA clustering algorithm starts forming new cluster.

Let $\mathrm{R}_{\mathscr{M}_i}(\mathbb{D})$ be the relevance of miRNAs $\mathscr{M}_i \in \mathbb{C}$ with respect to class label $\mathbb{D}$. The relevance uses information about the class labels and is thus a criterion for supervised clustering. The supervised clustering algorithm starts with a single miRNA $\mathscr{M}_i$ that has the highest relevance value with respect to class labels. An initial cluster $\mathbb{V}_i$ is then formed by selecting the set of miRNAs $\{\mathscr{M}_j\}$ from the whole set $\mathbb{C}$ considering the miRNA $\mathscr{M}_i$ as the representative of cluster $\mathbb{V}_i$, where

$$\mathbb{V}_i = \{\mathscr{M}_j | \psi(\mathscr{M}_i,\mathscr{M}_j) \geq \delta; \mathscr{M}_j \neq \mathscr{M}_i \in \mathbb{C}\}. \tag{22}$$

Hence, the cluster $\mathbb{V}_i$ represents the set of miRNAs of $\mathbb{C}$ those have the supervised similarity values with the miRNA $\mathscr{M}_i$ greater than a pre-defined threshold value δ. The cluster $\mathbb{V}_i$ is the coarse cluster corresponding to the miRNA $\mathscr{M}_i$, while the threshold δ is termed as the radius of cluster $\mathbb{V}_i$.

Once the initial cluster $\mathbb{V}_i$ is formed, the cluster representative is refined by adding other miRNAs to the cluster. By searching among the miRNAs of cluster $\mathbb{V}_i$, the current cluster representative is merged and averaged with one single miRNA such that the augmented cluster representative $\bar{\mathscr{M}}_i$ increases the relevance value. The merging process is repeated until the relevance value can no longer be improved. Instead of averaging all miRNAs of $\mathbb{V}_i$, the augmented attribute $\bar{\mathscr{M}}_i$ is computed by considering a subset of attributes/miRNAs $\bar{\mathbb{V}}_i \subset \mathbb{V}_i$ those increase the relevance value of cluster representative $\bar{\mathscr{M}}_i$. The set of attributes/miRNAs $\bar{\mathbb{V}}_i$ represents the finer cluster of the attribute $\mathscr{M}_i$. While the generation of coarse cluster reduces the redundancy among miRNAs of the set $\mathbb{C}$, that of finer cluster increases the relevance with respect to class labels. After generating the augmented cluster representative $\bar{\mathscr{M}}_i$ from the finer cluster $\bar{\mathbb{V}}_i$, the process is repeated to find more clusters and augmented cluster representatives by discarding the set of miRNAs $\bar{\mathbb{V}}_i$ from the whole set $\mathbb{C}$.

The main steps of the integrated miRNA clustering algorithm (ALGO1) are reported next.

• Let $\mathbb{C}$ be the set of miRNAs of the original data set, while $\mathbb{S}$ and $\bar{\mathbb{S}}$ are the set of actual and augmented attributes, respectively, selected by the miRNA clustering algorithm.

• Let $\mathbb{V}_i$ be the coarse cluster associated with the miRNA $\mathscr{M}_i$ and $\bar{\mathbb{V}}_i$, the finer cluster of $\mathscr{M}_i$, represents the set of miRNAs of $\mathbb{V}_i$ those are merged and averaged with the attribute $\mathscr{M}_i$ to generate the augmented cluster representative $\bar{\mathscr{M}}_i$.

(1) Initialize $\mathbb{C} \leftarrow \{\mathscr{M}_1,\ldots,\mathscr{M}_i,\ldots,\mathscr{M}_j,\ldots,\mathscr{M}_m\}$, $\mathbb{S} \leftarrow \varnothing$, and $\bar{\mathbb{S}} \leftarrow \varnothing$.

(2) Calculate the rough hypercuboid based relevance value $\mathrm{R}_{\mathscr{M}_i}(\mathbb{D})$ of each miRNA $\mathscr{M}_i \in \mathbb{C}$.

(3) Repeat the following nine steps (steps 4 to 12) until $\mathbb{C} = \varnothing$ or the desired number of attributes are selected.

(4) Select miRNA $\mathscr{M}_i$ from $\mathbb{C}$ as the representative of cluster $\mathbb{V}_i$ that has highest rough hypercuboid based relevance value. In effect, $\mathscr{M}_i \in \mathbb{S}$, $\mathscr{M}_i \in \mathbb{V}_i$, $\bar{\mathscr{M}}_i \in \bar{\mathbb{V}}$, and $\mathbb{C} = \mathbb{C}\backslash\mathscr{M}_i$.

(5) Generate coarse cluster $\mathbb{V}_i$ from the set of existing attributes/miRNAs of $\mathbb{C}$ satisfying the following condition:

$$\mathbb{V}_i = \{\mathscr{M}_j | \psi(\mathscr{M}_i, \mathscr{M}_j) \geq \delta; \mathscr{M}_j \neq \mathscr{M}_i \in \mathbb{C}\}.$$

(6) Initialize $\bar{\mathscr{M}}_i \leftarrow \mathscr{M}_i$.

(7) Repeat following four steps (steps 8 to 11) for each miRNA $\mathscr{M}_j \in \mathbb{V}_i$.

(8) Compute two augmented cluster representatives by averaging $\mathscr{M}_j$ and its complement with the attributes of $\bar{\mathbb{V}}_i$ as follows:

$$\bar{\mathscr{M}}^{+}_{i+j} = \frac{1}{|\bar{\mathbb{V}}_i| + 1}\left\{\sum_{\mathscr{M}_k \in \bar{\mathbb{V}}_i} \mathscr{M}_k + \mathscr{M}_j\right\} \quad (23)$$

$$\bar{\mathscr{M}}^{-}_{i+j} = \frac{1}{|\bar{\mathbb{V}}_i| + 1}\left\{\sum_{\mathscr{M}_k \in \bar{\mathbb{V}}_i} \mathscr{M}_k - \mathscr{M}_j\right\} \quad (24)$$

(9) The augmented cluster representative $\bar{\mathscr{M}}_{i+j}$ after averaging $\mathscr{M}_j$ or its complement with $\bar{\mathbb{V}}_i$ is as follows:

$$\bar{\mathscr{M}}_{i+j} = \begin{cases} \bar{\mathscr{M}}^{+}_{i+j} & \text{if } \mathrm{R}_{\bar{\mathscr{M}}^{+}_{i+j}}(\mathbb{D}) \geq \mathrm{R}_{\bar{\mathscr{M}}^{-}_{i+j}}(\mathbb{D}) \\ \bar{\mathscr{M}}^{-}_{i+j} & \text{otherwise} \end{cases}. \quad (25)$$

(10) The augmented cluster representative $\bar{\mathscr{M}}_i$ of cluster $\mathbb{V}_i$ is $\bar{\mathscr{M}}_{i+j}$ if $\mathrm{R}_{\bar{\mathscr{M}}_{i+j}}(\mathbb{D}) \geq \mathrm{R}_{\bar{\mathscr{M}}_i}(\mathbb{D})$, otherwise $\bar{\mathscr{M}}_i$ remains unchanged.

(11) Select attribute $\mathscr{M}_j$ or its complement as a member of the finer cluster $\bar{\mathbb{V}}_i$ of attribute $\mathscr{M}_i$ if $\mathrm{R}_{\bar{\mathscr{M}}_{i+j}}(\mathbb{D}) \geq \mathrm{R}_{\bar{\mathscr{M}}_i}(\mathbb{D})$.

(12) In effect, $\bar{\mathscr{M}}_i \in \bar{\mathbb{S}}$ and $\mathbb{C} = \mathbb{C}\backslash\bar{\mathbb{V}}_i$.

(13) Stop.

In this regard, it can be shown that as the number of desired clusters is constant and sufficiently small compared to the total number of attributes m, the proposed clustering algorithm has an overall $\mathcal{O}(m)$ time complexity.

3.4 *B*.632+ error rate

In order to minimize the variability and biasedness of the results obtained by the proposed clustering algorithm, the so-called *B*.632+ bootstrap approach[65] is used, which is defined as follows:

$$B.632+ = (1 - \tilde{\omega})\mathrm{AE} + \tilde{\omega}B1 \quad (26)$$

where AE denotes the proportion of the original training samples misclassified, termed as apparent error rate, and $B1$ is the bootstrap error, defined as follows:

$$B1 = \frac{1}{n}\sum_{j=1}^{n}\left(\frac{\sum_{k=1}^{M} I_{jk}Q_{jk}}{\sum_{k=1}^{M} I_{jk}}\right) \quad (27)$$

where n is the number of original samples and M is the number of bootstrap samples. If the sample x_j is not contained in the kth bootstrap sample, then $I_{jk} = 1$, otherwise 0. Similarly, if x_j is misclassified, $Q_{jk} = 1$, otherwise 0. The weight parameter $\tilde{\omega}$ is given by

$$\tilde{\omega} = \frac{0.632}{1 - 0.368r}; \quad (28)$$

$$\text{where } r = \frac{B1 - \mathrm{AE}}{\gamma - \mathrm{AE}}; \quad (29)$$

$$\text{and } \gamma = \sum_{i=1}^{c} p_i(1 - q_i); \quad (30)$$

where c is the number of classes, p_i is the proportion of the samples from the ith class, and q_i is the proportion of them assigned to the ith class. Also, γ is termed as the no-information error rate that would apply if the distribution of the class-membership label of the sample x_j did not depend on its feature vector.

3.5 Support vector machine

In the current study, the support vector machine (SVM)[67] is used to evaluate the performance of the proposed rough hypercuboid based supervised clustering algorithm as well as several other algorithms. The SVM is a margin classifier that draws an optimal hyperplane in the feature vector space; this defines a boundary that maximizes the margin between data 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 and radial basis function kernels are used. The source code of the SVM has been downloaded from Library for Support Vector Machines (www.csie.ntu.edu.tw/~cjlin/libsvm/).

To compute different types of error rates obtained using the SVM, bootstrap approach is performed on each miRNA expression data set. For each training set, a set of averaged differential attributes/miRNAs is first generated, and then the SVM is trained with the selected averaged miRNA expression values. After the training, the information of averaged miRNAs those were selected for the training set is used to generate test set and then the class label of the test sample is predicted using the SVM. For each data set, fifty top-ranked miRNA clusters are selected for the analysis.

In order to calculate the *B*.632+ error rate, apparent error (AE) is first calculated. This error is obtained when the same original data set is used to train and test a classifier. After that, the $B1$ error is computed from M bootstrap samples. Finally, the no-information error (γ) is calculated by randomly perturbing the class label of a given data set. The mutated data set is used to find groups of miRNA and the coherent expression of miRNAs is used to build the SVM. Then, the trained SVM is used to classify the original data set. The error generated by this procedure is known as γ error rate. Finally, the *B*.632+ error rate is computed based on the AE, $B1$ error, and γ error using (26).

Together with the rough hypercuboid based supervised miRNA clustering algorithm described in the Section 3.3 and *B*.632+ bootstrap approach, a set of differentially expressed miRNAs are selected. Following are the main steps of the proposed method:

(1) Generate M bootstrap samples from the original data set.

(2) Each M bootstrap sample is comprised of training set and test set. On training set, apply ALGO1 and use the selected miRNA information to design SVM classifier. Check the effectiveness of the classifier in terms of error by using test set for that bootstrap sample. Similarly, calculate error for rest of the bootstrap samples. Based on these errors, $B1$ error rate can be calculated using (27).

(3) Compute apparent error (AE) by applying ALGO1 on the original data set. Build the SVM classifier using the selected miRNA information. Then check the performance of the classifier by using same data set that is used to train the classifier.

(4) No information error or γ error can be calculated using (30) by randomly permuting the class labels of the original data set. Here, on the mutated data set apply ALGO1 then build classifier based on the output of ALGO1 then use the original data set for checking the effectiveness of the classifier.

(5) Finally, using (26) calculate *B*.632+ error rate for the entire data set.

4 Experimental results

The performance of the proposed rough hypercuboid equivalence partition matrix based supervised miRNA clustering (RH-SAC) method is extensively studied and compared with that of some existing feature selection and clustering algorithms. The algorithms compared are mutual information based InfoGain[68] and minimum redundancy-maximum relevance (mRMR) algorithm,[69] method proposed by Golub *et al.*,[70] rough set based maximum relevance-maximum significance (RSMRMS) algorithm,[61,71] μHEM,[62] and fuzzy-rough supervised attribute clustering algorithm (FR-SAC).[55] The maximum number of features selected by the proposed supervised miRNA clustering algorithm is 50. Two types of kernels are used for SVM, linear and RBF kernels. Default parameters of libSVM software for RBF kernels are used. The source code of the proposed RH-SAC algorithm, written in C language, is available at https://sites.google.com/site/sushmitapaulsite/Home/RH–SACPackage.tar.gz?attredirects=0&d=1. All the algorithms are run in Ubuntu 12.04 LTS having machine configuration Intel Core i7-2600 CPU @ 3.40 GHz × 8, and 16 GB RAM.

4.1 miRNA expression data sets used

In this paper, publicly available five miRNA expression data sets are used to establish the effectiveness of the proposed approach. The five miRNA expression data sets with accession number GSE17846, GSE21036, GSE24709, GSE28700, and GSE31408 are downloaded from Gene Expression Omnibus (www.ncbi.nlm.nih.gov/geo/).

GSE17846. This data set represents the analysis of miRNA profiling in peripheral blood samples of multiple sclerosis and in the blood of normal donors. It contains 864 miRNAs, 41 samples, and 2 classes.[72]

GSE21036. This data set contains miRNA expression profiles of 218 prostate tumors with primary or metastatic prostate cancer with a median of 5 years clinical follow-up. The number of miRNAs and samples are 373 and 141, respectively.[73]

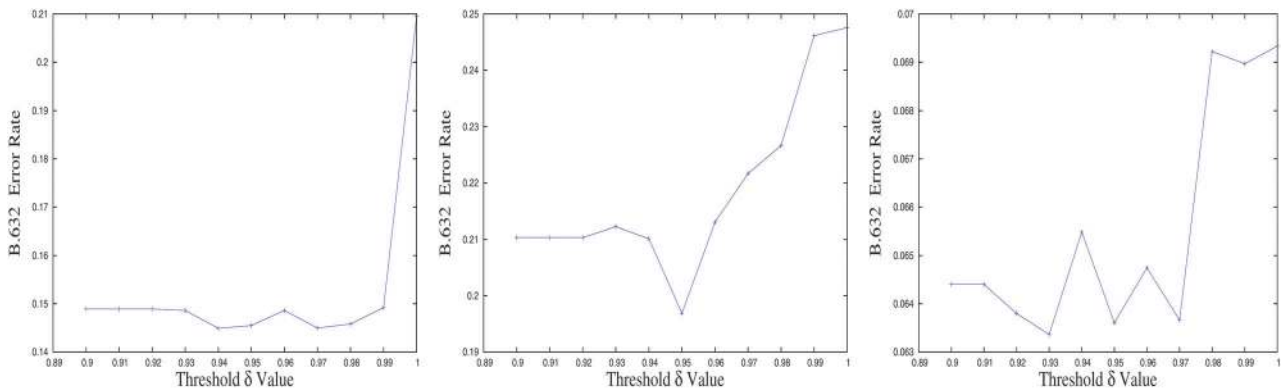
GSE24709. It analyzes peripheral miRNA blood profiles of patients with lung diseases. The miRNA expression profiling has been done for patients with lung cancer, chronic obstructive pulmonary disease, and normal controls. It contains 863 miRNAs, 71 samples, and 3 classes.

GSE28700. This data set contains expression profiles of miRNAs from 22 paired gastric cancer and normal tissues. It contains total 44 samples and 470 miRNAs. The samples are grouped into 2 classes.[74]

GSE31408. It analyzes miRNA expression profile of cutaneous T-cell lymphomas and benign inflammation of skin. It consists of 705 miRNAs, 148 samples, and 2 classes.[75]

4.2 Optimal value of δ parameter

The threshold δ in (22) plays an important role in the performance of the proposed supervised miRNA clustering algorithm.



(a) GSE24709 (b) GSE28700 (c) GSE31408

Fig. 1 Variation of *B*.632+ error rate for different values of threshold δ.

Table 1 Optimum values of δ^* for different miRNA data sets

Microarray data sets	Optimum δ^*	Number of miRNAs
GSE17846	0.99	31
GSE21036	1.00	49
GSE24709	0.94	43
GSE28700	0.95	43
GSE31408	0.93	23

It controls the size of a cluster. Hence, it has direct influence on the performance of the proposed algorithm. Higher the value of δ sparse the cluster becomes. To find the optimal value, the proposed algorithm is implemented on five data sets. The value for which the $B.632+$ error rate is minimum is considered to be the optimum δ value for the corresponding data set. Here, the results are reported with respect to the errors generated by the linear kernels of SVM.

Fig. 1 represents the variation of $B.632+$ error rate with respect to the value of δ. The value of δ is varied from 0.90 to 1.00. From the figure, it is seen that as the δ value increases the $B.632+$ error rate decreases and achieves a minimum value and then the error rate again increases with the increment of the δ value. Hence, the optimum values of δ for five miRNA data sets are calculated using the following relation:

$$\delta^* = \arg\min_{\delta}\{B.632 + \text{error}\}. \quad (31)$$

However, in data set GSE31408, the $B.632+$ error rate fluctuates for δ range 0.94 to 0.97. It suggests that the proposed supervised clustering algorithm gets stuck into local minima of the search space for this range. Table 1 represents the optimal δ^* values for all the miRNA data sets. The table also presents the number of miRNAs at which optimal δ^* value is obtained for miRNA data sets.

4.3 Different types of errors

This section describes about the different types of errors generated by the SVM classifier. The importance of $B.632+$ error over apparent error (AE), gamma error (γ), and bootstrap

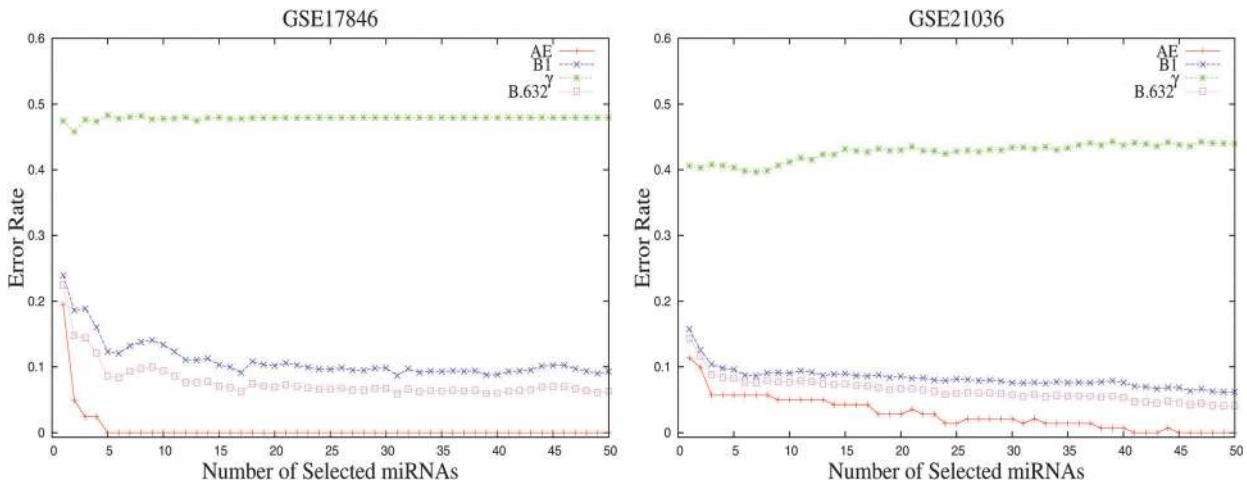


Fig. 2 Different error rates of the proposed algorithm on GSE17846 and GSE21036 sets obtained using the SVM averaged over 50 random splits.

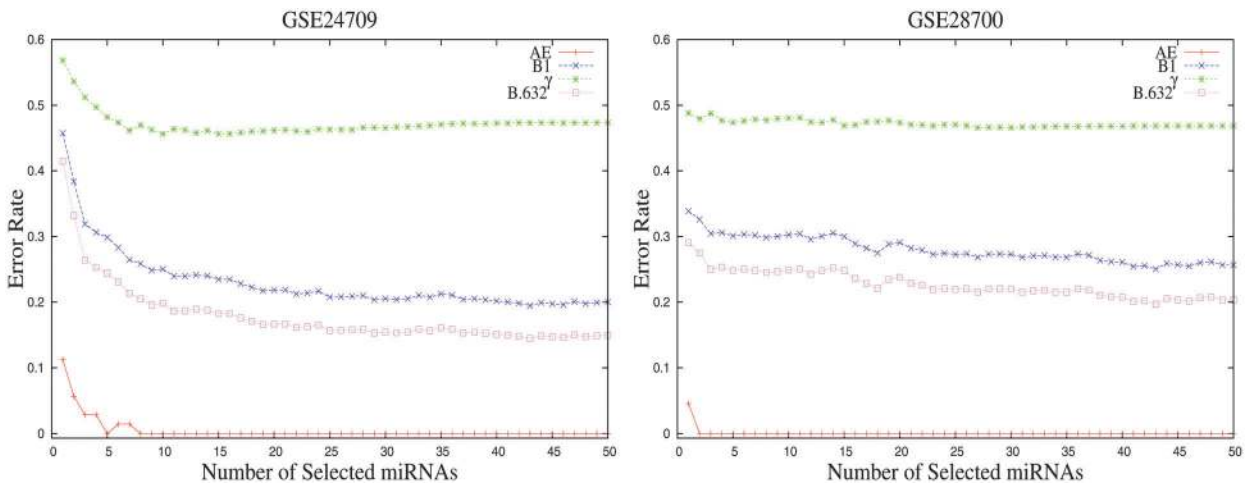


Fig. 3 Different error rates of the proposed algorithm on GSE24709 and GSE28700 sets obtained using the SVM averaged over 50 random splits.

Table 2 Comparative analysis of different types of errors for proposed method

	AE		$B1$ error		γ error		$B.632+$ error	
Microarray data sets	Error	miRNAs	Error	miRNAs	Error	miRNAs	Error	miRNAs
GSE17846	0.000	5	0.087	31	0.458	2	0.059	31
GSE21036	0.000	41	0.062	49	0.397	7	0.041	49
GSE24709	0.000	5	0.195	43	0.456	10	0.145	43
GSE28700	0.000	2	0.250	43	0.466	27	0.197	43
GSE31408	0.000	4	0.092	23	0.389	4	0.063	23

($B1$) error is also established. All the errors are calculated using the SVM for the proposed method. The results are presented for the optimum values of δ. Fig. 2 and 3 represent different types of errors obtained for five different data sets. From the figures, it is seen that the γ error rate is higher than any other type of errors for each data set, while $B1$ error is lower than the γ error rate but higher than the $B.632+$ error and AE. The weighted average of $B1$ error and AE leads to $B.632+$ error rate lower than the $B1$ error but higher than AE.

Table 2 represents minimum values of different types of errors and corresponding number of miRNAs at which the error is obtained for each miRNA data set. From the table, it is seen that the $B.632+$ estimator rectifies the upward bias of $B1$ error and downward bias of AE.

4.4 Effectiveness of Kernel functions

Table 3 represents the comparative performance of the two kernel functions of the SVM. In this work, linear kernels and radial basis function kernels are used to compute different errors. However, the errors generated by these two kernels based SVM are comparable. From the table it is seen that out of 34 cases the linear kernel based SVM generates low $B.632+$ error rate in 19 cases, while the RBF kernel based SVM performs better in 15 cases. Sign test has been used to check the statistical significance of linear kernels over RBF kernels. Although, the P-value of the linear kernels over RBF kernels is low but not significant. The obtained P-value is 0.3642. The proposed supervised miRNA clustering algorithm selects set of miRNAs that is highly differentially expressed. Hence, the classifier that is designed by using output of the proposed method generates low error rate irrespective of kernels. It is also seen from the table that irrespective of any kernels the proposed supervised attribute clustering algorithm generates low $B.632+$ error rate compare to other algorithms except in one case each for linear kernels and RBF kernels, respectively.

4.5 Comparative performance analysis

In this section, comparative performance analysis of the proposed supervised miRNA clustering algorithm has been shown with respect to some popular feature selection and supervised attribute clustering algorithms.

Table 4 represents the different types of errors obtained by different methods at their optimal parameters. It also contains the number of miRNAs at which the corresponding lowest error rate is obtained by each method. From the table, it is seen that almost all the algorithms generate AE equal to zero. However, the InfoGain and FR-SAC generate non-zero AE in one case each. On the other hand, the RSMRMS generates non-zero AE in 4 cases. From the table, it is seen that the proposed supervised miRNA clustering algorithm generates $B.632+$ error rate lower than any other method except in one case. Only in one case, the μ-HEM algorithm generates better result than the proposed method. Fig. 4–6 compare $B.632+$ error rates generated by different methods. From the figures, it is seen that in most of the cases the proposed rough hypercuboid based supervised miRNA clustering algorithm performs better than any existing method.

4.6 Pathway enrichment analysis of obtained miRNAs

In this section, biological importance of the obtained miRNAs using proposed supervised miRNA clustering algorithm is described. The miRNAs, which are selected by the proposed

Table 3 Comparative analysis of different Kernels

		Golub	InfoGain	mRMR	RSMRMS	μ-HEM	FR-SAC	RH-SAC
GSE17846	Linear	0.0809	0.0630	0.0690	0.0640	0.0590	0.1803	0.0588
	RBF	0.0686	0.0616	0.0593	0.0597	0.0567	0.1933	0.0616
GSE21036	Linear	0.0466	0.0490	0.0430	0.0750	0.0390	0.0530	0.0410
	RBF	0.0558	0.0592	0.0514	0.0779	0.0473	0.0702	0.0421
GSE24709	Linear	*	0.2030	0.1910	0.3660	0.1800	0.2396	0.1449
	RBF	*	0.1893	0.1680	0.3617	0.1792	0.2565	0.1472
GSE28700	Linear	0.2482	0.2710	0.2850	0.2850	0.2570	0.2888	0.1969
	RBF	0.2698	0.2998	0.3096	0.3501	0.2570	0.3301	0.2365
GSE31408	Linear	0.0689	0.0770	0.0740	0.0770	0.0670	0.0793	0.0634
	RBF	0.0692	0.0725	0.0639	0.0738	0.0664	0.0670	0.0621

Table 4 Comparative performance analysis of different algorithms

		Apparent error		*B*1 error		γ error		*B*.632+ error	
Microarray data sets	Algorithms/methods	Error	miRNAs	Error	miRNAs	Error	miRNAs	Error	miRNAs
GSE17846	Golub	0.0000	6	0.1165	48	0.4795	48	0.0809	48
	InfoGain	0.0000	7	0.0930	37	0.4799	37	0.0630	37
	mRMR	0.0000	3	0.1010	48	0.4798	48	0.0690	48
	RSMRMS	0.0000	2	0.0930	39	0.4792	39	0.0640	39
	μ-HEM	0.0000	2	0.0870	49	0.4790	49	0.0590	49
	FR-SAC	0.0000	2	0.2340	47	0.4659	18	0.1803	47
	RH-SAC	0.0000	5	0.0870	31	0.4580	2	*0.0588*	31
GSE21036	Golub	0.0000	35	0.0694	48	0.4370	39	0.0466	48
	InfoGain	0.0000	39	0.0730	50	0.4452	44	0.0490	50
	mRMR	0.0000	19	0.0640	49	0.4400	50	0.0430	49
	RSMRMS	0.0500	5	0.0890	5	0.4173	5	0.0750	5
	μ-HEM	0.0000	42	0.0580	47	0.4440	47	*0.0390*	47
	FR-SAC	0.0000	41	0.0785	50	0.4020	1	0.0530	50
	RH-SAC	0.0000	41	0.0620	49	0.3970	7	0.0410	49
GSE24709	InfoGain	0.0000	26	0.2570	45	0.4747	46	0.2030	45
	mRMR	0.0000	24	0.2450	50	0.4737	50	0.1910	50
	RSMRMS	0.1410	11	0.4020	11	0.5250	2	0.3660	11
	μ-HEM	0.0000	20	0.2340	49	0.4750	49	0.1800	49
	FR-SAC	0.0000	49	0.2931	50	0.4755	50	0.2396	50
	RH-SAC	0.0000	5	0.1950	43	0.4560	10	*0.1449*	43
GSE28700	Golub	0.0000	27	0.3004	27	0.4736	3	0.2482	27
	InfoGain	0.0000	35	0.3090	8	0.4678	8	0.2710	21
	mRMR	0.0000	21	0.3330	49	0.4728	7	0.2850	49
	RSMRMS	0.0230	34	0.3310	19	0.4715	15	0.2850	19
	μ-HEM	0.0000	25	0.3060	4	0.5000	4	0.2570	4
	FR-SAC	0.0000	24	0.3362	50	0.4650	43	0.2888	50
	RH-SAC	0.0000	2	0.2500	43	0.4660	27	*0.1969*	43
GSE31408	Golub	0.0000	36	0.0734	1	0.4374	1	0.0689	1
	InfoGain	0.0070	20	0.0900	9	0.4213	1	0.0770	27
	mRMR	0.0000	37	0.0940	6	0.4253	1	0.0740	6
	RSMRMS	0.0610	2	0.0860	6	0.4218	2	0.0770	6
	μ-HEM	0.0000	44	0.0980	2	0.4520	50	0.0670	50
	FR-SAC	0.0068	44	0.1071	11	0.4181	13	0.0793	50
	RH-SAC	0.0000	4	0.0920	23	0.3890	4	*0.0634*	23

method in all the 50 bootstrap samples, were used for further analysis. The association of these miRNAs with different biological pathways was determined.

The DIANA-miRPath v2.0[76] interface has been used to identify the miRNA–pathway relationship. The server performs an enrichment analysis of miRNA gene targets in KEGG pathways.[77]

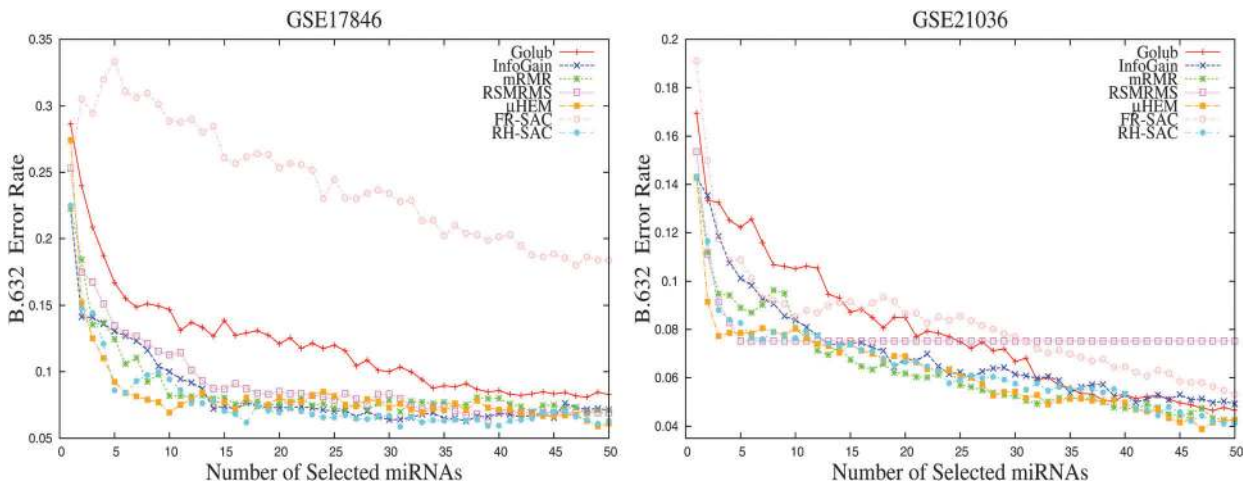


Fig. 4 *B*.632+ errors of the SVM obtained using different methods on GSE17846 and GSE21036 data sets averaged over 50 random splits.

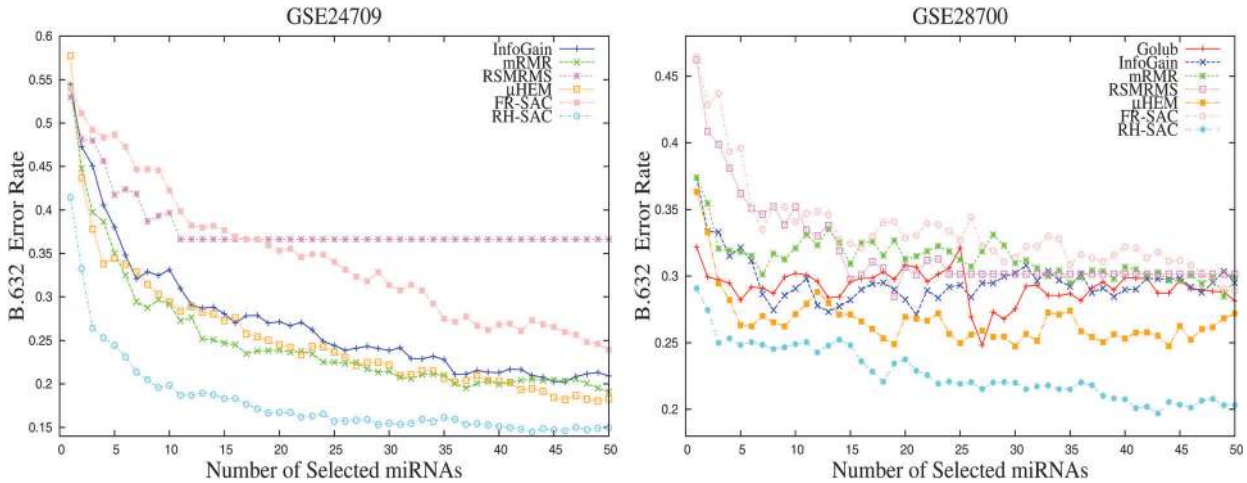


Fig. 5 *B*.632+ errors of the SVM obtained using different methods on GSE24709 and GSE28700 data sets averaged over 50 random splits.

The tool first identifies the target genes of the uploaded miRNAs. A user can select gene targets from Tarbase database. It is a database that contains miRNA gene target information that are experimentally validated. Another target prediction algorithm, namely, DIANA-microT-CDS[78] can also be applied to predict target genes of miRNAs. The algorithm utilizes one-tailed Fisher's exact test to detect enriched pathways with targets of specific miRNAs. In this server, the false discovery rate (FDR) method is also available as a correction for multiple hypothesis testing. Subsequently, a *P*-value is calculated for both types of test. The nearer the *P*-value is to zero, more statistically significant that pathway is. This tool also generates a heatmap. Heat maps are graphical representation of data, where values in a matrix are represented as colors. In this tool, hierarchical clustering is implemented on miRNAs and pathways in order to construct the heat map. The generated heatmaps can be used to see the miRNA–network relationships.

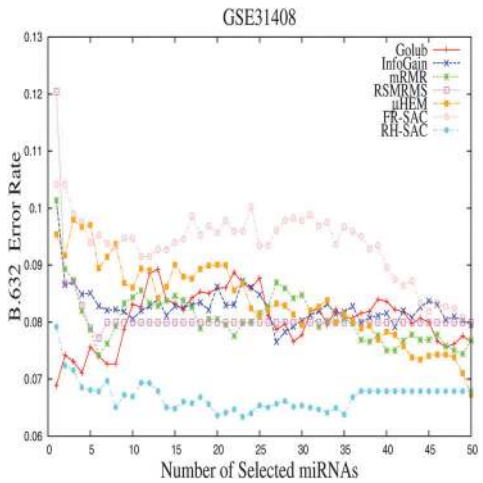


Fig. 6 *B*.632+ errors of the SVM obtained using different methods on GSE31408 data set averaged over 50 random splits.

The DIANA-miRPath v2.0 has been applied on the selected miRNAs of three miRNA data sets, namely, GSE17846, GSE21036, and GSE28700. Those pathways are selected whose *P*-value is lower than 0.05. Fig. 7–9 represent the heatmap of the miRNA-pathways, which are found to be statistically significant. The darker colors represent that the miRNA is associated with the pathway more significantly. Fig. 7 represents the heat map for data set GSE17846. This data contains miRNA profiling from total blood of multiple sclerosis and control samples. From the figure, it is seen that the miRNAs selected by the

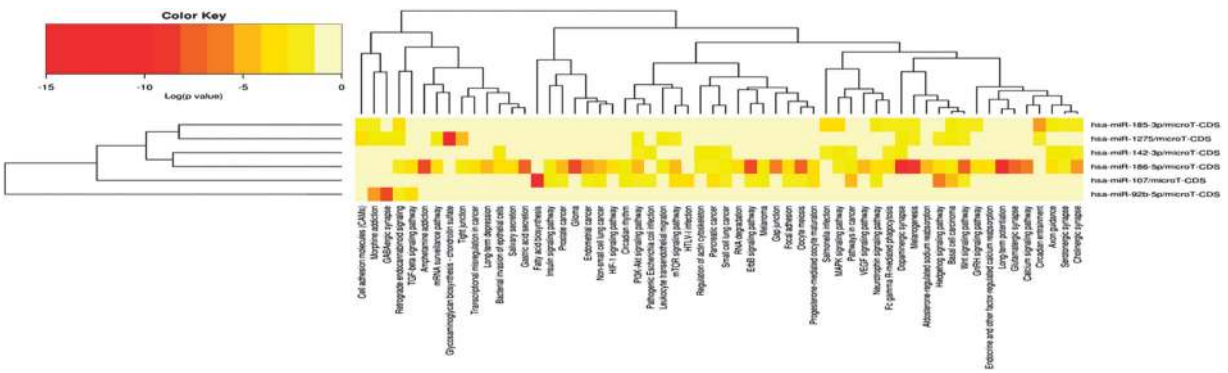


Fig. 7 miRNAs *versus* pathways heat map for GSE17846.

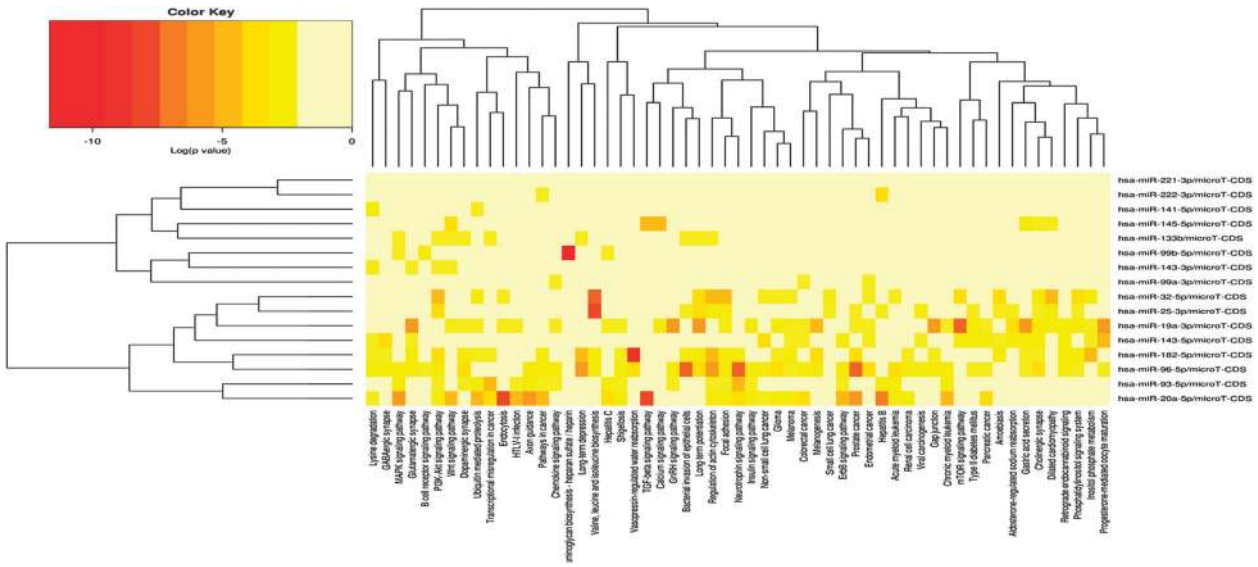


Fig. 8 miRNAs *versus* pathways heat map for GSE21036.

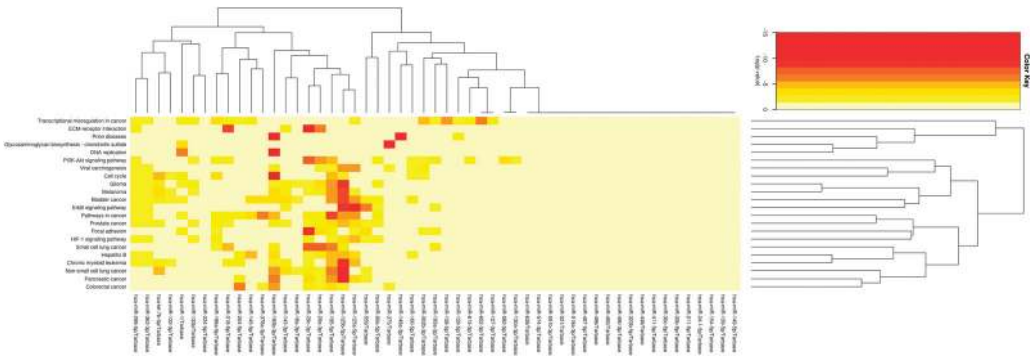


Fig. 9 miRNAs *versus* pathways heat map for GSE21036.

proposed method are statistically related with 29 pathways. Multiple sclerosis is an autoimmune disorder and from the Fig. 7 it is seen that around 7 pathways are significant and they are related to autoimmune disorder. They are Cell adhesion molecules (CAMs), TGF-beta signaling pathway, PI3K-Akt signaling pathway, Leukocyte transendothelial migration, MAPK signaling pathway, Fc gamma R-mediated phagocytosis, and calcium signaling pathway. Other than these pathways, other pathways are also found to be statistically significant and they are cancer, morphine addiction, GABAergic synapse, glioma prostate cancer, endometrial cancer, and ErbB signaling pathway. On the other hand, around 48 pathways–miRNAs relationship are found to be statistically significant for GSE21036 data set. This data set is generated using metastatic prostate cancer samples and normal adjacent benign prostate. From Fig. 8, it is seen that the proposed method is able to select those miRNAs that are associated with prostate cancer. In addition to that it is also able to identify other significant pathways. They are progestrone-mediated oocyte maturation, inositol phosphate metabolism, mTOR signaling pathway, hepatitis B, melanogenesis, neurotrophin signaling pathway, TGF-beta signaling pathway, chemokine signaling pathway, which are found significantly associated with the selected miRNAs. Similarly, several significant miRNA–pathway relations are obtained using the DIANA-miRPath tool for the data set GSE28700. In this data set, expression profiles of microRNAs in gastric cancer are stored. From Fig. 9, it is clear several cancer related pathways are found to be significant using the proposed method. From the figure, it is seen that total 22 pathways are found to be significant and they are colorectal cancer, pancreatic cancer, non-small cell lung cancer, chronic myeloid leukemia, hepatitis B, small cell lung cancer, HIF-1 signaling pathway, focal adhesion, prostate cancer, pathways in cancer, ErbB signaling pathway, bladder cancer, melanoma, glioma, cell cycle, viral carcinogenesis, PI3K-Akt signaling pathway, DNA replication, glycosaminoglycan biosynthesis-chondroitin sulfate, prion diseases, ECM–receptor interaction, and transcriptional misregulation in cancer.

5 Conclusion

The paper presents a new rough hypercuboid based supervised miRNA clustering algorithm. It uses the concept of rough

hypercuboid equivalence partition matrix for calculating similarity between two miRNAs and thus improves the performance of the method. The rough hypercuboid based similarity measure uses the information of class label for calculating similarity between two miRNAs and hence, makes it a supervised measure. The proposed method fetches cluster of miRNAs whose collective expression is strongly associated with the class label. The effectiveness of the proposed rough hypercuboid based supervised miRNA clustering algorithm is shown and compared with other existing methods on five miRNA expression data sets. The selected miRNAs are also found to be significantly associated with different important pathways that are related to the data set. The new method is capable of identifying effective miRNAs that may contribute to revealing underlying etiology of a disease, providing a useful tool for exploratory analysis of miRNA data.

Acknowledgements

The authors want to acknowledge Dr Pradipta Maji of Indian Statistical Institute, Kolkata, India for his valuable suggestions. This work was supported by the German Federal Ministry of Education and Research (BMBF) as part of the projects eBio:-miRSys [0316175A to JV]. Julio Vera is funded by the Erlangen University Hospital (ELAN funds, 14-07-22-1-Vera-Gonzlez) and the German Research Foundation (DFG) through the project SPP 1757/1 (VE 642/1-1 to JV). Sushmita Paul is funded by the Erlangen University Hospital.

References

1 L. He and G. J. Hannon, *Nat. Rev. Genet.*, 2004, **5**, 522–531.
2 B. D. Harfe, *Curr. Opin. Genet. Dev.*, 2005, **15**, 410–415.
3 N. Bushati and S. M. Cohen, *Annu. Rev. Cell Dev. Biol.*, 2007, **23**, 175–205.
4 J. Krol, I. Loedige and W. Filipowicz, *Nat. Rev. Genet.*, 2010, **11**, 597–610.
5 G. A. Calin, M. Ferracin, A. Cimmino, G. Di Leva, M. Shimizu, S. E. Wojcik, M. V. Iorio, R. Visone, N. I. Sever, M. Fabbri, R. Iuliano, T. Palumbo, F. Pichiorri, C. Roldo, R. Garzon, C. Sevignani, L. Rassenti, H. Alder, S. Volinia, C. G. Liu, T. J. Kipps, M. Negrini and C. M. Croce, *N. Engl. J. Med.*, 2005, **353**, 1793–1801.
6 I. Alvarez-Garcia and E. A. Miska, *Development*, 2005, **132**, 4653–4662.
7 P. S. Mitchell, R. K. Parkin, E. M. Kroh, B. R. Fritz, S. K. Wyman, E. L. Pogosova-Agadjanyan, A. Peterson, J. Noteboom, K. C. O'Briant, A. Allen, D. W. Lin, N. Urban, C. W. Drescher, B. S. Knudsen, D. L. Stirewalt, R. Gentleman, R. L. Vessella, P. S. Nelson, D. B. Martin and M. Tewari, *Proc. Natl. Acad. Sci. U. S. A.*, 2008, **105**, 10513–10518.
8 K. Beezhold, V. Castranova and F. Chen, *Mol. Cancer*, 2010, **9**, 134.
9 Y. Altuvia, P. Landgraf, G. Lithwick, N. Elefant, S. Pfeffer, A. Aravin, M. J. Brownstein, T. Tuschl and H. Margalit, *Nucleic Acids Res.*, 2005, **33**, 2697–2706.
10 S. Paul and P. Maji, *Mol. BioSyst.*, 2014, **10**, 1509–1523.
11 W. C. Chan, M. R. Ho, S. C. Li, K. W. Tsai, C. H. Lai, C. N. Hsu and W. C. Lin, *Genomics*, 2012, **100**, 141–148.
12 J. Wang, M. Haubrock, K. M. Cao, X. Hua, C. Y. Zhang, E. Wingender and J. Li, *BMC Syst. Biol.*, 2011, **5**, 199.
13 M. Lagos-Quintana, R. Rauhut, J. Meyer, A. Borkhardt and T. Tuschl, *RNA*, 2003, **9**, 175–179.
14 E. Lai, P. Tomancak, R. Williams and G. Rubin, *Genome Biol.*, 2003, **4**, R42.
15 E. Thatcher, J. Bond, I. Paydar and J. Patton, *BMC Genomics*, 2008, **9**, 253.
16 A. F. Olena and J. G. Patton, *J. Cell. Physiol.*, 2010, **222**, 540–545.
17 S. Baskerville and D. P. Bartel, *RNA*, 2005, **11**, 241–247.
18 X. Cai, C. H. Hagedorn and B. R. Cullen, *RNA*, 2004, **10**, 1957–1966.
19 Y. Lee, M. Kim, J. Han, K. H. Yeom, S. Lee, S. H. Baek and V. N. Kim, *EMBO J.*, 2004, **23**, 4051–4060.
20 J. Lu, G. Getz, E. A. Miska, E. A. Saavedra, J. Lamb, D. Peck, A. S. Cordero, B. L. Ebert, R. H. Mak, A. A. Ferrando, J. R. Downing, T. Jacks, H. R. Horvitz and T. R. Golub, *Nat. Lett.*, 2005, **435**, 834–838.
21 C. Blenkiron, L. D. Goldstein, N. P. Thorne, I. Spiteri, S. F. Chin, M. J. Dunning, N. L. Barbosa-Morais, A. E. Teschendorff, A. R. Green, I. O. Ellis, S. Tavare, C. Caldas and E. A. Miska, *Genome Biol.*, 2007, **8**, 214.1–214.16.
22 Y. Chen and R. L. Stallings, *Cancer Res.*, 2007, **67**, 976–983.
23 J. Guo, Y. Miao, B. Xiao, R. Huan, Z. Jiang, D. Meng and Y. Wang, *J. Gastroenterol. Hepatol.*, 2009, **24**, 652–657.
24 M. G. Schrauder, R. Strick, R. Schulz-Wendtland, P. L. Strissel, L. Kahmann, C. R. Loehberg, M. P. Lux, S. M. Jud, A. Hartmann, A. Hein, C. M. Bayer, M. R. Bani, S. Richter, B. R. Adamietz, E. Wenkel, C. Rauh, M. W. Beckmann and P. A. Fasching, *PLoS One*, 2012, **7**, 1–9.
25 H. Zhao, J. Shen, L. Medico, D. Wang, C. B. Ambrosone and S. Liu, *PLoS One*, 2010, **5**, 1–12.
26 S. Arora, A. R. Ranade, N. L. Tran, S. Nasser, S. Sridhar, R. L. Korn, J. T. D. Ross, H. Dhruv, K. M. Foss, Z. Sibenaller, T. Ryken, M. B. Gotway, S. Kim and G. J. Weiss, *Int. J. Cancer*, 2011, **129**, 2621–2631.
27 A. D. McIver, P. East, C. A. Mein, J. B. Cazier, G. Molloy, T. Chaplin, T. A. Lister, B. D. Young and S. Debernardi, *PLoS One*, 2008, **3**, 1–8.
28 C. Wang, S. Yang, G. Sun, X. Tang, S. Lu, O. Neyrolles and Q. Gao, *PLoS One*, 2011, **6**, 1–11.
29 M. Zhu, M. Yi, C. H. Kim, C. Deng, Y. Li, D. Medina, R. M. Stephens and J. E. Green, *Genome Biol.*, 2011, **12**, 1–16.
30 M. V. Iorio, R. Visone, G. D. Leva, V. Donati, F. Petrocca, P. Casalini, C. Taccioli, S. Volinia, C. G. Liu, H. Alder, G. A. Calin, S. Menard and C. M. Croce, *Cancer Res.*, 2007, **67**, 8699–8707.
31 S. Li, X. Chen, H. Zhang, X. Liang, Y. Xiang, C. Yu, K. Zen, Y. Li and C. Y. Zhang, *J. Lipid Res.*, 2009, **50**, 1756–1765.
32 S. Nasser, A. R. Ranade, S. Sridhart, L. Haney, R. L. Korn, M. B. Gotway, G. J. Weiss and S. Kim, *Proceedings of IEEE International Conference on Bioinformatics and Biomedicine*, 2010, pp. 246–251.

33 F. J. Ortega, J. M. Moreno-Navarrete, G. Pardo, M. Sabater, M. Hummel, A. Ferrer, J. I. Rodriguez-Hermosa, B. Ruiz, W. Ricart, B. Peral and J. M. F. Real, *PLoS One*, 2010, **5**, 1–9.
34 P. M. Pereira, J. P. Marques, A. R. Soares, L. Carreto and M. A. S. Santos, *PLoS One*, 2010, **5**, 1–12.
35 M. Raponi, L. Dossey, T. Jatkoe, X. Wu, G. Chen, H. Fan and D. G. Beer, *Cancer Res.*, 2009, **69**, 5776–5783.
36 C. Xue, F. Li, T. He, G.-P. Liu, Y. Li and X. Zhang, *BMC Bioinf.*, 2005, **6**, 310.
37 B. Liu, L. Fang, J. Chen, F. Liu and X. Wang, *Mol. BioSyst.*, 2015, **11**, 1194–1204.
38 B. Liu, L. Fang, F. Liu, X. Wang, J. Chen and K. C. Chou, *PLoS One*, 2015, **10**, e0121501.
39 B. Liu, L. Fang, F. Liu, X. Wang and K. C. Chou, *J. Biomol. Struct. Dyn.*, 2015, **3**, 1–13.
40 E. Enerly, I. Steinfeld, K. Kleivi, S. K. Leivonen, M. R. Aure, H. G. Russnes, J. A. Ronneberg, H. Johnsen, R. Navon, E. Rodland, R. Makela, B. Naume, M. Perala, O. Kallioniemi, V. N. Kristensen, Z. Yakhini and A. L. B. Dale, *PLoS One*, 2011, **6**, e16915.
41 R. Bargaje, M. Hariharan, V. Scaria and B. Pillai, *RNA*, 2010, **16**, 16–25.
42 L. J. Heyer, S. Kruglyak and S. Yooseph, *Genome Res.*, 1999, **9**, 1106–1115.
43 A. Ben-Dor, R. Shamir and Z. Yakhini, *J. Comput. Biol.*, 1999, **6**, 281–297.
44 E. Hartuv and R. Shamir, *Inform. Process. Lett.*, 2000, **76**, 175–181.
45 R. Shamir and R. Sharan, *Proceedings of the 8th International Conference on Intelligent Systems for Molecular Biology*, 2000, pp. 307–331.
46 E. P. Xing and R. M. Karp, *Bioinformatics*, 2001, **17**, 306–315.
47 C. Fraley and A. E. Raftery, *Comput. J.*, 1998, **41**, 578–588.
48 D. Ghosh and A. M. Chinnaiyan, *Bioinformatics*, 2002, **18**, 275–286.
49 G. J. McLachlan, R. W. Bean and D. Peel, *Bioinformatics*, 2002, **18**, 413–422.
50 K. Y. Yeung, C. Fraley, A. Murua, A. E. Raftery and W. L. Ruzz, *Bioinformatics*, 2001, **17**, 977–987.
51 P. Maji and S. Paul, *Fundamenta Informaticae*, 2013, **124**, 153–174.
52 D. Jiang, J. Pei and A. Zhang, *Proceedings of the 3rd IEEE International Symposium on Bioinformatics and Bioengineering*, 2003, pp. 393–400.
53 M. Dettling and P. Buhlmann, *Genome Biol.*, 2002, **3**, 1–15.
54 P. Maji, *IEEE Transactions on Knowledge and Data Engineering*, 2012, **24**, 127–140.
55 P. Maji, *IEEE Transactions on System, Man and Cybernetics, Part B: Cybernetics*, 2011, **41**, 222–233.
56 T. Hastie, R. Tibshirani, D. Botstein and P. Brown, *Genome Biol.*, 2001, **1**, 1–12.
57 Z. Pawlak, *Rough Sets: Theoretical Aspects of Resoning About Data*, Kluwer, Dordrecht, The Netherlands, 1991.
58 P. Maji and S. K. Pal, *Rough-Fuzzy Pattern Recognition: Applications in Bioinformatics and Medical Imaging*, Wiley-IEEE Computer Society Press, New Jersey, 2012.
59 P. Maji and S. K. Pal, *IEEE Transactions on System, Man and Cybernetics, Part B: Cybernetics*, 2007, **37**, 1529–1540.
60 P. Maji and S. K. Pal, *IEEE Transactions on System, Man, and Cybernetics, Part B: Cybernetics*, 2010, **40**, 741–752.
61 P. Maji and S. Paul, *International Journal of Approximate Reasoning*, 2011, **52**, 408–426.
62 S. Paul and P. Maji, *BMC Bioinf.*, 2013, **14**, 266.
63 P. Maji and S. Paul, *IEEE/ACM Trans. Comput. Biol. Bioinf.*, 2013, **10**, 286–299.
64 P. Maji, *IEEE Transactions on Knowledge and Data Engineering*, 2014, **26**, 16–29.
65 B. Efron and R. Tibshirani, *J. Am. Stat. Assoc.*, 1997, **92**, 548–560.
66 J.-M. Wei, S.-Q. Wang and X.-J. Yuan, *IEEE Transactions on Knowledge and Data Engineering*, 2010, **22**, 381–391.
67 V. Vapnik, *The Nature of Statistical Learning Theory*, Springer-Verlag, New York, 1995.
68 J. R. Quinlan, *C4.5: Programs for Machine Learning*, Morgan Kaufmann, CA, 1993.
69 C. Ding and H. Peng, *J. Bioinf. Comput. Biol.*, 2005, **3**, 185–205.
70 T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield and E. S. Lander, *Science*, 1999, **286**, 531–537.
71 S. Paul and P. Maji, *Int. J. Nanomed.*, 2013, 1–17.
72 A. Keller, P. Leidinger, J. Lange, A. Borries, H. Schroers, M. Scheffler, H.-P. Lenhof, K. Ruprecht and E. Meese, *PLoS One*, 2009, **4**, e7440.
73 B. S. Taylor, N. Schultz, H. Hieronymus, A. Gopalan, Y. Xiao, B. S. Carver, V. K. Arora, P. Kaushik, E. Cerami, B. Reva, Y. Antipin, N. Mitsiades, T. Landers, I. Dolgalev, J. E. Major, M. Wilson, N. D. Socci, A. E. Lash, A. Heguy, J. A. Eastham, H. I. Scher, V. E. Reuter, P. T. Scardino, C. Sander, C. L. Sawyers and W. L. Gerald, *Cancer Cell*, 2010, **18**, 11–22.
74 C.-W. Tseng, C.-C. Lin, C.-N. Chen, H.-C. Huang and H.-F. Juan, *BMC Syst. Biol.*, 2011, **5**, 99.
75 U. Ralfkiaer, P. H. Hagedorn, N. Bangsgaard, M. B. Lovendorf, C. B. Ahler, L. Svensson, K. L. Kopp, M. T. Vennegaard, B. Lauenborg, J. R. Zibert, T. Krejsgaard, C. M. Bonefeld, R. Sokilde, L. M. Gjerdrum, T. Labuda, A.-M. Mathiesen, K. Gronbaek, M. A. Wasik, M. Sokolowska-Wojdylo, C. Queille-Roussel, R. Gniadecki, E. Ralfkiaer, C. Geisler, T. Litman, A. Woetmann, C. Glue, M. A. Ropke, L. Skov and N. Odum, *Blood*, 2011, **118**, 5891–5900.
76 I. S. Vlachos, N. Kostoulas, T. Vergoulis, G. Georgakilas, M. Reczko, M. Maragkakis, M. D. Paraskevopoulou, K. Prionidis, T. Dalamagas and A. G. Hatzigeorgiou, *Nucleic Acids Res.*, 2012, **40**, W498–W504.
77 M. Kanehisa, S. Goto, Y. Sato, M. Kawashima, M. Furumichi and M. Tanabe, *Nucleic Acids Res.*, 2014, **42**, D199–D205.
78 M. Reczko, M. Maragkakis, P. Alexiou, I. Grosse and A. G. Hatzigeorgiou, *Bioinformatics*, 2012, 1–8.