Supervised Cluster Analysis of miRNA Expression Data Using Rough Hypercuboid Partition Matrix

Sushmita Paul(✉) and Julio Vera

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

Abstract. The microRNAs are small, endogenous non-coding RNAs found in plants and animals, which suppresses the expression of genes post-transcriptionally. 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 background, a new rough hypercuboid based supervised similarity measure is proposed that is integrated with the supervised attribute clustering to find groups of miRNAs whose coherent expression can classify samples. The proposed method directly incorporates the information of sample categories into the miRNA clustering process, generating a supervised clustering algorithm for miRNAs. The effectiveness of the rough hypercuboid based algorithm, along with a comparison with other related algorithms, is demonstrated on three miRNA microarray expression data sets using the $B.632+$ bootstrap error rate of support vector machine. The association of the miRNA clusters to various biological pathways are also shown by doing pathway enrichment analysis.

Keywords: MicroRNA · Co-expressed miRNAs · Clustering · Rough sets

1 Introduction

Micro RNAs/miRNAs are a class of short approximately 22-nucleotide non-coding RNAs found in many plants and animals. They inhibit the expression of mRNA expression post-transcriptionally. It has been shown by [1] that the miRNAs on a genome tend to present in a cluster. Large scale surveys [2] have established the fact that miRNAs have tendency to present in clusters. Existence of co-expressed miRNAs is also demonstrated using expression profiling analysis in [3]. 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. In [4,15], different approaches are introduced to discover

M. Kryszkiewicz et al. (Eds.): PReMI 2015, LNCS 9124, pp. 482–494, 2015.
DOI: 10.1007/978-3-319-19941-2_46

miRNA cluster patterns. 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.

Several unsupervised clustering techniques like hierarchical clustering algorithms [8] and self organizing maps [2] are used to cluster a miRNA expression data. However, the groups of miRNAs discovered by these unsupervised clustering algorithms are not potential enough to do tissue classification [5], 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 [5,10,11]. In [5], 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 [10]. In [11], a fuzzy-rough supervised gene clustering algorithm is described. The algorithm 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.

However, 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 [16] provides a mathematical framework to capture uncertainties associated with human cognition process. In [11,13,14], rough sets have been successfully used to analyze a microarray expression data.

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 [12,14] and supervised attribute clustering algorithm [11]. 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 developed for calculating similarity between two miRNAs. Based upon the similarity values, the miRNAs are grouped into cluster. The new supervised clustering algorithm divides the miRNA expression data in 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 [7] 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.

2 Rough Hypercuboid Based Supervised Attribute Clustering

In this paper, a new algorithm is developed based on rough hypercuboid equivalence partition matrix. Every clustering algorithm need a distance or similarity measure to group objects. Accordingly, a new rough hypercuboid based similarity measure is proposed. The concept of rough hypercuboid was presented in [20], while that of rough hypercuboid equivalence partition matrix was proposed in [12,14]. It has also been successfully applied for feature/gene/miRNA selection in [12,14]. The relevance of a cluster is calculated using rough hypercuboid equivalence partition matrix based dependency measure. The proposed rough hypercuboid based supervised similarity measure is integrated into the supervised attribute clustering algorithm developed by Maji [11]. Prior to describe about the new supervised attribute clustering algorithm, next the concept of rough hypercuboid equivalence partition matrix is described.

2.1 Rough Hypercuboid Equivalence Partition Matrix

Let $\mathbb{U} = \{s_1, \cdots, s_i, \cdots, s_n\}$ be the set of n objects or samples and $\mathbb{C} = \{\mathscr{M}_1, \cdots, \cdots, \mathscr{M}_m\}$ denotes the set of m attributes or miRNAs of a given microarray data set. Let $\mathbb{D}$ be the set of class labels or sample categories of n samples.

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 or miRNA $\mathscr{M}_k \in \mathbb{C}$. If $\mathbb{U}/\mathscr{M}_k = \{\mu_1, \cdots, \mu_i, \cdots, \mu_c\}$ 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 $\mathbb{H}(\mathscr{M}_k) = [\mathrm{h}_{ij}(\mathscr{M}_k)]$. The matrix $\mathbb{H}(\mathscr{M}_k)$ is denoted by

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

$$\text{where } \mathrm{h}_{ij}(\mathscr{M}_k) = \begin{cases} 1 \text{ if } \mathrm{L}_i \leq x_j(\mathscr{M}_k) \leq \mathrm{U}_i \\ 0 \text{ otherwise.} \end{cases} \tag{2}$$

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.

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. The $c \times n$ matrix $\mathbb{H}(\mathscr{M}_k)$ is termed as hypercuboid equivalence partition matrix of the condition attribute or miRNA $\mathscr{M}_k$. 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 class β_i satisfying following two conditions:

$$1 \leq \sum_{j=1}^{n} \mathrm{h}_{ij}(\mathscr{M}_k) \leq n, \forall i; \quad 1 \leq \sum_{i=1}^{c} \mathrm{h}_{ij}(\mathscr{M}_k) \leq c, \forall j. \tag{3}$$

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.

Using the concept of hypercuboid equivalence partition matrix, the misclassified objects of boundary region present in the implicit hypercuboids can be identified based on the confusion vector defined next

$$\mathbb{V}(\mathscr{M}_k) = [\mathrm{v}_1(\mathscr{M}_k), \cdots, \cdots, \mathrm{v}_n(\mathscr{M}_k)]; \text{ where } \mathrm{v}_j(\mathscr{M}_k) = \min\{1, \sum_{i=1}^{c} \mathrm{h}_{ij}(\mathscr{M}_k) - 1\}. \tag{4}$$

In rough sets 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 $\mathrm{v}_j(\mathscr{M}_k) = 0$. 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 $\mathrm{v}_j(\mathscr{M}_k) = 1$. 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:

$$\underline{M}(\beta_i) = \{s_j | \ \mathrm{h}_{ij}(\mathscr{M}_k) = 1 \text{ and } \mathrm{v}_j(\mathscr{M}_k) = 0\}; \quad \overline{M}(\beta_i) = \{s_j | \ \mathrm{h}_{ij}(\mathscr{M}_k) = 1\}; \tag{5}$$

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

$$BN_M(\beta_i) = \{s_j | \ \mathrm{h}_{ij}(\mathscr{M}_k) = 1 \text{ and } \mathrm{v}_j(\mathscr{M}_k) = 1\}. \tag{6}$$

Dependency. The dependency between condition attribute $\mathscr{M}_k$ and decision attribute $\mathbb{D}$ can be defined as follows:

$$\gamma_{\mathscr{M}_k}(\mathbb{D}) = \frac{1}{n}\sum_{i=1}^{c}\sum_{j=1}^{n} \mathrm{h}_{ij}(\mathscr{M}_k) \cap [1 - \mathrm{v}_j(\mathscr{M}_k)]; \text{ that is, } \gamma_{\mathscr{M}_k}(\mathbb{D}) = 1 - \frac{1}{n}\sum_{j=1}^{n} \mathrm{v}_j(\mathscr{M}_k), \quad (7)$$

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. The resultant hypercuboid equivalence partition matrix $\mathbb{H}(\{\mathscr{M}_k, \mathscr{M}_l\})$ of size $c \times n$ can be computed from $\mathbb{H}(\mathscr{M}_k)$ and $\mathbb{H}(\mathscr{M}_l)$ as follows:

$$\mathbb{H}(\{\mathscr{M}_k, \mathscr{M}_l\}) = \mathbb{H}(\mathscr{M}_k) \cap \mathbb{H}(\mathscr{M}_l); \text{ where } \mathrm{h}_{ij}(\{\mathscr{M}_k, \mathscr{M}_l\}) = \mathrm{h}_{ij}(\mathscr{M}_k) \cap \mathrm{h}_{ij}(\mathscr{M}_l). \quad (8)$$

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

$$\sigma_{\mathbb{M}}(\mathbb{D}, \mathscr{M}_k) = \frac{1}{n}\sum_{j=1}^{n} \left[\mathrm{v}_j(\mathbb{M} - \{\mathscr{M}_k\}) - \mathrm{v}_j(\mathbb{M})\right]; \quad (9)$$

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.

2.2 Rough Hypercuboid Based Supervised Similarity Measure

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

Let $\mathbb{C} = \{\mathscr{M}_1, \cdots, \mathscr{M}_i, \cdots, \mathscr{M}_j, \cdots, \mathscr{M}_{\mathcal{D}}\}$ denotes the set of $\mathcal{D}$ 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:

$$\mathrm{R}_{\mathscr{M}_i}(\mathbb{D}) = \gamma_{\mathscr{M}_i}(\mathbb{D}) \quad (10)$$

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 (7).

At first, the distance between two miRNAs $\mathscr{M}_i$ and $\mathscr{M}_j$ is calculated using rough hypercuboid based approach. Then the non-linear transformation of the distance 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 (9) is used to compute similarity between two miRNAs and it is defined next.

Definition 1. *The rough hypercuboid based similarity measure between two attributes or miRNAs $\mathscr{M}_i$ and $\mathscr{M}_j$ is defined as follows:*

$$\psi(\mathscr{M}_i, \mathscr{M}_j) = \frac{1}{\sqrt{\kappa^2 + 1}}; \quad \text{where} \quad \kappa = \left\{ \frac{\sigma_{\mathscr{M}_i}(\mathbb{D}, \mathscr{M}_j) + \sigma_{\mathscr{M}_j}(\mathbb{D}, \mathscr{M}_i)}{2} \right\} \tag{11}$$

Hence, the supervised similarity measure $\psi(\mathscr{M}_i, \mathcal{A}_j)$ 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 miRNAs $\mathscr{M}_i$ and $\mathscr{M}_j$.

2.3 Supervised miRNA Clustering Algorithm

In this work the proposed rough hypercuboid based similarity measure is incorporated into the Fuzzy-Rough Supervised Attribute Clustering Algorithm [11]. In the proposed method a new rough hypercuboid based similarity measure is developed to calculate similarity between two miRNAs. Whereas, in [11] 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.

Let $\mathbb{C}$ represents 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 $\mathbb{V}_i$ is 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 $\bar{\mathscr{M}}_i$. The main steps of the integrated miRNA clustering algorithm are reported next.

1. Initialize $\mathbb{C} \leftarrow \{\mathscr{M}_1, \cdots, \mathscr{M}_i, \cdots, \mathscr{M}_j, \cdots, \mathscr{M}_{\mathcal{D}}\}$, $\mathbb{S} \leftarrow \emptyset$, and $\bar{\mathbb{S}} \leftarrow \emptyset$.
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 $\mathbb{C} = \emptyset$ 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, $\mathscr{M}_i \in \mathbb{S}$, $\mathscr{M}_i \in \mathbb{V}_i$, $\mathscr{M}_i \in \bar{\mathbb{V}}_i$, and $\mathbb{C} = \mathbb{C} \setminus \mathscr{M}_i$.

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

$$\mathbb{V}_i = \{\mathscr{M}_j | \psi(\mathscr{M}_i, \mathscr{M}_j) \geq \delta; \mathscr{M}_j \neq \mathscr{M}_i \in \mathbb{C}\}. \tag{12}$$

6. Initialize $\bar{\mathscr{M}}_i \leftarrow \mathscr{M}_i$.
7. Repeat following four steps (steps 8–11) 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:

$$\bar{\mathscr{M}}^{+}_{i+j} = \frac{1}{|\bar{\mathbb{V}}_i| + 1} \left\{ \sum_{\mathscr{M}_k \in \bar{\mathbb{V}}_i} \mathscr{M}_k + \mathscr{M}_j \right\}; \bar{\mathscr{M}}^{-}_{i+j} = \frac{1}{|\bar{\mathbb{V}}_i| + 1} \left\{ \sum_{\mathscr{M}_k \in \bar{\mathbb{V}}_i} \mathscr{M}_k - \mathscr{M}_j \right\} \tag{13}$$

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:

$$\bar{\mathscr{M}}_{i+j} = \begin{cases} \bar{\mathscr{M}}^{+}_{i+j} \text{ if } \mathrm{R}_{\bar{\mathscr{M}}^{+}_{i+j}}(\mathbb{D}) \geq \mathrm{R}_{\bar{\mathscr{M}}^{-}_{i+j}}(\mathbb{D}) \\ \bar{\mathscr{M}}^{-}_{i+j} \text{ otherwise.} \end{cases} \tag{14}$$

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 $\mathbb{C} = \mathbb{C} \setminus \bar{\mathbb{V}}_i$.
13. Stop.

3 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 on three miRNA expression data sets GSE17846, GSE21036, and GSE28700. The algorithms compared are mutual information based Info-Gain [17] and minimum redundancy-maximum relevance (mRMR) algorithm [6], method proposed by Golub et al. [9], rough set based maximum relevance-maximum significance (RSMRMS) algorithm [13], μHEM [14], fuzzy-rough supervised attribute clustering algorithm (FR-SAC) [11]. The error rate of support vector machine (SVM) [18] is used to evaluate the performance of different algorithms. To compute the error rate of SVM, bootstrap approach (B.632+ error rate) [7] is performed on each miRNA expression data set. For each training set, a set of differential miRNA groups is first generated, and then SVM is trained with the selected coherent miRNAs. After the training, the information of 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 classifier. The maximum number of features selected by the new integrated supervised miRNA clustering algorithm are 50.

3.1 Optimal Value of δ Parameter

The threshold δ in (12) plays an important role in the performance of the proposed supervised miRNA clustering algorithm. It controls the size of a cluster. Hence, it has direct influence in the performance of the proposed algorithm. Higher the value of δ sparse the cluster becomes. To find the optimal value of δ parameter the proposed algorithm is implemented on three 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.The value of δ is varied from 0.90 to 1.00. Hence, the optimum value of δ for three miRNA data sets are calculated using the following relation:

$$\delta^{\star} = \arg\min_{\delta}\{B.632 + \text{error}\}. \tag{15}$$

The optimum values of δ^* obtained using (15) are 0.99, 1.00, 0.95 for GSE17846, GSE21036, and GSE28700 data sets, respectively. The number of miRNAs at which optimal δ^* value is obtained for miRNA data sets are 31, 49, and 43 for GSE17846, GSE21036, and GSE28700 data sets, respectively.

3.2 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 ($B1$) 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. Figure 1 represents different types of errors obtained for three different data sets. From the figure 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 average of $B1$ error and AE leads to $B.632+$ error rate lower than the $B1$ error but higher than AE. Table 1 represents minimum values of different types of errors and corresponding number of miRNAs at which the error is obtained for each miRNA data sets. From the table it is seen that the $B.632+$ estimator rectifies the upward bias of $B1$ error and downward bias of AE.

Table 1. Comparative analysis of different types of errors for proposed method

Microarray data sets	AE		B1 Error		γ Error		B.632+ Error	
	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
GSE28700	0.000	2	0.250	43	0.466	27	0.197	43

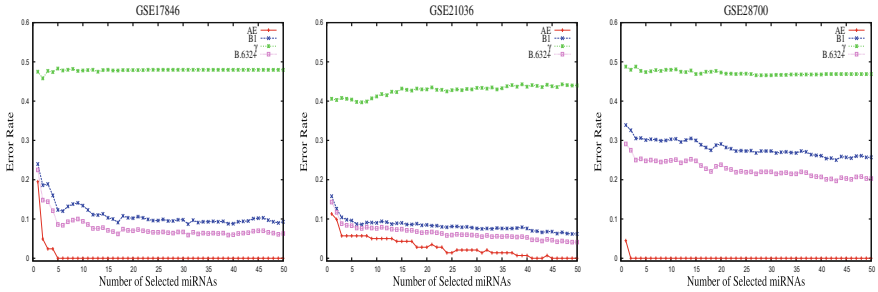


Fig. 1. Different error rates of the proposed algorithm on different data sets obtained using the SVM averaged over 50 random splits

3.3 Comparative Performance Analysis

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

Table 2 represents the different types of error 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 the almost all the algorithms generate AE equal to zero. However, the RSMRMS generates non zero AE in 2 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 miRNA selection algorithm generates better result than the proposed method.

3.4 Pathway Enrichment Analysis of Obtained miRNAs

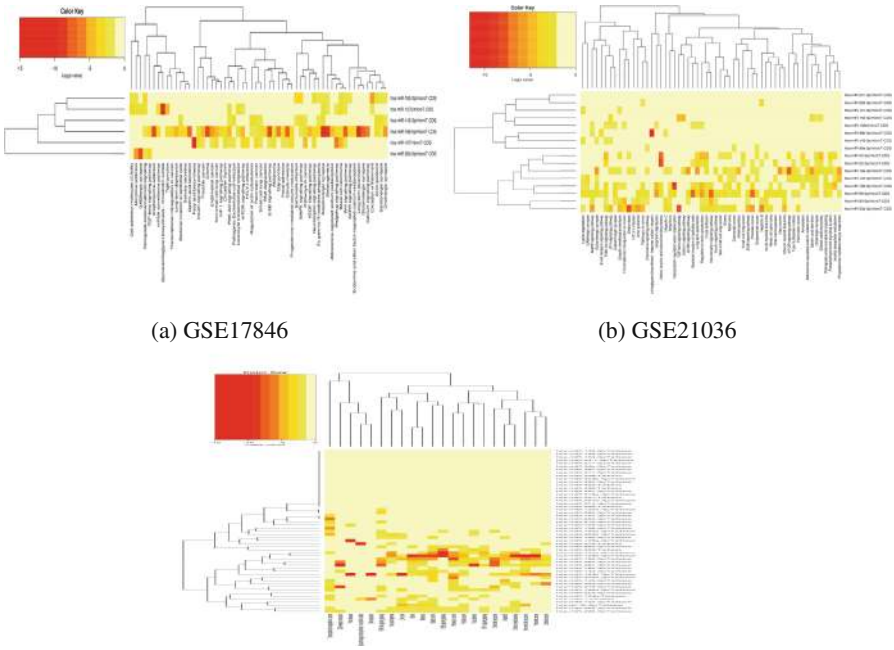
In this section biological importance of the obtained miRNAs using proposed supervised miRNA clustering algorithm is described. Those miRNAs which are selected by the proposed method in all the 50 bootstrap samples were used for further analysis. The association of those miRNAs with different biological pathways were determined. The DIANA-miRPath v2.0 [19] interface has been used to identify the miRNA-pathway relationship. The server performs an enrichment analysis of miRNA gene targets in KEGG pathways. The tool first identifies the target genes of the uploaded miRNAs.

The DIANA-miRPath v2.0 has been applied on the selected miRNAs of miRNA data sets. Those pathways are selected whose P-value is lower than 0.05. The miRNA-pathway relation is represented by a heatmap. Figure 2 represents 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. In data set GSE17846 the miRNA profiling of total blood of multiple sclerosis and control samples is performed. From the figure it is

Table 2. Comparative performance analysis of different algorithms

Microarray data sets	Algorithms/ Methods	Apparent Error		$B1$ Error		γ Error		$B.632+$ Error	
		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
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

seen the miRNAs selected by the proposed method are statistically related with 29 pathways. Multiple Sclerosis is a autoimmune disorder and from the Fig. 2 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. 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. 2 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 like Progestrone-mediated oocyte maturation, Inositol phosphate metabolism, mTOR signaling pathway, and so forth. 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. 2 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 few of them 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 cencer, Pathways in cancer.



(a) GSE17846 (b) GSE21036

(c) GSE28700

Fig. 2. miRNAs versus pathways heat map for different miRNA data sets

4 Conclusion

The paper presents a new rough hypercuboid based supervised similarity measure that is incorporated into the supervised miRNA clustering algorithm. It uses the concept of rough hypercuboid 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 expressions are 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 three 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.

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 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 through the project SPP 1757/1 (VE 642/1-1 to JV). Sushmita Paul is funded by the Erlangen University Hospital.

References

1. Altuvia, Y., Landgraf, P., Lithwick, G., Elefant, N., Pfeffer, S., Aravin, A., Brownstein, M.J., Tuschl, T., Margalit, H.: Clustering and conservation patterns of human microRNAs. Nucleic Acids Res. **33**, 2697–2706 (2005)
2. Bargaje, R., Hariharan, M., Scaria, V., Pillai, B.: Consensus miRNA expression profiles derived from interplatform normalization of microarray data. RNA **16**, 16–25 (2010)
3. Baskerville, S., Bartel, D.P.: Microarray profiling of microRNAs reveals frequent coexpression with neighboring miRNAs and host genes. RNA **11**, 241–247 (2005)
4. Chan, W.C., Ho, M.R., Li, S.C., Tsai, K.W., Lai, C.H., Hsu, C.N., Lin, W.C.: MetaMirClust: discovery of miRNA cluster patterns using a data-mining approach. Genomics **100**(3), 141–148 (2012)
5. Dettling, M., Buhlmann, P.: Supervised clustering of genes. Genome Biol. **3**(12), 1–15 (2002)
6. Ding, C., Peng, H.: Minimum redundancy feature selection from microarray gene expression data. J. Bioinform. Comput. Biol. **3**(2), 185–205 (2005)
7. Efron, B., Tibshirani, R.: Improvements on cross-validation: the.632+ bootstrap method. J. Am. Stat. Assoc. **92**(438), 548–560 (1997)
8. Enerly, E., Steinfeld, I., Kleivi, K., Leivonen, S.K., Aure, M.R., Russnes, H.G., Ronneberg, J.A., Johnsen, H., Navon, R., Rodland, E., Makela, R., Naume, B., Perala, M., Kallioniemi, O., Kristensen, V.N., Yakhini, Z., Dale, A.L.B.: miRNA-mRNA integrated analysis reveals roles for miRNAs in primary breast tumors. PLoS ONE **6**(2), e16915 (2011)
9. Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A., Bloomfield, C.D., Lander, E.S.: Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science **286**(5439), 531–537 (1999)
10. Hastie, T., Tibshirani, R., Botstein, D., Brown, P.: Supervised harvesting of expression trees. Genome Biol. **1**, 1–12 (2001)
11. Maji, P.: Fuzzy-rough supervised attribute clustering algorithm and classification of microarray data. IEEE Trans. Syst. Man Cybern. B Cybern. **41**(1), 222–233 (2011)
12. Maji, P.: A rough hypercuboid approach for feature selection in approximation spaces. IEEE Trans. Knowl. Data Eng. **26**(1), 16–29 (2014)
13. Maji, P., Paul, S.: Rough set based maximum relevance-maximum significance criterion and gene selection from microarray data. Int. J. Approximate Reasoning **52**(3), 408–426 (2011)
14. Paul, S., Maji, P.: μHEM for identification of differentially expressed miRNAs using hypercuboid equivalence partition matrix. BMC Bioinform. **14**(1), 266 (2013)
15. Paul, S., Maji, P.: City block distance and rough-fuzzy clustering for identification of co-expressed MicroRNAs. Mol. BioSyst. **10**(6), 1509–1523 (2014)
16. Pawlak, Z.: Rough Sets: Theoretical Aspects of Resoning About Data. Kluwer, Dordrecht (1991)
17. Quinlan, J.R.: C4.5: Programs for Machine Learning. Morgan Kaufmann, San Francisco (1993)

18. Vapnik, V.: The Nature of Statistical Learning Theory. Information Science and Statistics. Springer, New York (1995)
19. Vlachos, I.S., Kostoulas, N., Vergoulis, T., Georgakilas, G., Reczko, M., Maragkakis, M., Paraskevopoulou, M.D., Prionidis, K., Dalamagas, T., Hatzigeorgiou, A.G.: DIANA miRPath v. 2.0: investigating the combinatorial effect of microRNAs in pathways. Nucleic Acids Res. **40**(W1), W498–W504 (2012)
20. Wei, J.-M., Wang, S.-Q., Yuan, X.-J.: Ensemble rough hypercuboid approach for classifying cancers. IEEE Trans. Knowl. Data Eng. **22**(3), 381–391 (2010)