International Journal of Approximate Reasoning 52 (2011) 408–426

Contents lists available at ScienceDirect

International Journal of Approximate Reasoning

journal homepage: www.elsevier.com/locate/ijar

Rough set based maximum relevance-maximum significance criterion and gene selection from microarray data

Pradipta Maji *, Sushmita Paul

Machine Intelligence Unit, Indian Statistical Institute, 203, Barrackpore Trunk Road, Kolkata 700 108, India

ARTICLE INFO

Article history:
Received 26 May 2010
Revised 17 August 2010
Accepted 24 September 2010
Available online 20 October 2010

Keywords:
Microarray analysis
Gene selection
Rough sets
Feature selection
Classification

ABSTRACT

Among the large amount of genes presented in microarray gene expression data, only a small fraction of them is effective for performing a certain diagnostic test. In this regard, a new feature selection algorithm is presented based on rough set theory. It selects a set of genes from microarray data by maximizing the relevance and significance of the selected genes. A theoretical analysis is presented to justify the use of both relevance and significance criteria for selecting a reduced gene set with high predictive accuracy. The importance of rough set theory for computing both relevance and significance of the genes is also established. The performance of the proposed algorithm, along with a comparison with other related methods, is studied using the predictive accuracy of K-nearest neighbor rule and support vector machine on five cancer and two arthritis microarray data sets. Among seven data sets, the proposed algorithm attains 100% predictive accuracy for three cancer and two arthritis data sets, while the rough set based two existing algorithms attain this accuracy only for one cancer data set.

1. Introduction

Recent advancement and wide use of high-throughput technology are producing an explosion in using gene expression phenotype for identification and classification in a variety of diagnostic areas. An important application of gene expression data in functional genomics is to classify samples according to their gene expression profiles such as to classify cancer versus normal samples or to classify different types or subtypes of cancer [1,2].

A microarray gene expression data set can be represented by an expression table, $\mathcal{T} = \{w_{ij} | i = 1, \ldots, m, j = 1, \ldots, n\}$, where $w_{ij} \in \Re$ is the measured expression level of gene $\mathcal{A}_i$ in the jth sample, m and n represent the total number of genes and samples, respectively. Each row in the expression table corresponds to one particular gene and each column to a sample [1,2]. However, for most gene expression data, the number of training samples is still very small compared to the large number of genes involved in the experiments. The number of samples is likely to remain small for many areas of investigation, especially for human data, due to the difficulty of collecting and processing microarray samples [1]. When the number of genes is significantly greater than the number of samples, it is possible to find biologically relevant correlations of gene behavior with the sample categories [3,4].

However, among the large amount of genes, only a small fraction of them is effective for performing a certain task. Also, a small subset of genes is desirable in developing gene expression based diagnostic tools for delivering precise, reliable, and interpretable results. With the gene selection results, the cost of biological experiment and decision can be greatly reduced by analyzing only the marker genes. Hence, identifying a reduced set of most relevant and significant genes is the goal of gene selection. The small number of training samples and a large number of genes make gene selection a more relevant and

* Corresponding author. Tel.: +91 33 25753113; fax: +91 33 25788699.
E-mail addresses: pmaji@isical.ac.in (P. Maji), sushmita_t@isical.ac.in (S. Paul).

0888-613X/$ - see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.ijar.2010.09.006

challenging problem in gene expression based classification. This is an important problem in machine learning and referred to as feature selection [5–7].

Conventional methods of feature selection involve evaluating different feature subsets using some index and selecting the best among them. Depending on the way of computing the feature evaluation index, feature selection methods are generally divided into two broad categories: filter approach [5,6,8–11] and wrapper approach [5,7,12–14]. Unlike wrapper approach [5,7,12–15], in filter approach, the algorithms do not perform classification of the data in the process of feature evaluation. Before application of the actual learning algorithm, the best subset of features is selected in one pass by evaluating some predefined criteria, which are independent of the actual generalization performance of the learning machine. Hence, the filter approach is computationally less expensive and more general than that of wrapper approach. However, as the wrapper approach uses the learning machine as a black box, it generally outperforms the filter approach in the aspect of final predictive accuracy of the learning machine [5–15].

In feature selection process, an optimal feature subset is always relative to a certain criterion. In general, different criteria may lead to different optimal feature subsets. However, every criterion tries to measure the discriminating ability of a feature or a subset of features to distinguish different class labels. To measure the gene-class relevance, different statistical and information theoretic measures such as the F-test, t-test [8,9], entropy, information gain, mutual information [8,10], normalized mutual information [11], and f-information [16] are typically used, and the same or a different metric like mutual information, f-information, the L_1 distance, Euclidean distance, and Pearson's correlation coefficient [8,10,17] is employed to calculate the gene–gene redundancy. However, as the F-test, t-test, Euclidean distance, and Pearson's correlation depend on the actual gene expression values of the microarray data, they are very much sensitive to noise or outlier of the data set [8,10,17,18]. On the other hand, as information measures depend only on the probability distribution of a random variable rather than on its actual values, they are more effective to evaluate both gene-class relevance and gene–gene redundancy [10,11,19–21].

Rough set theory [22,23] is a new paradigm to deal with uncertainty, vagueness, and incompleteness. It has been applied to fuzzy rule extraction [24], reasoning with uncertainty, fuzzy modeling, feature selection [25–28], microarray data analysis [20,21,29,30], and so forth. It is proposed for indiscernibility in classification according to some similarity [22,31]. The rough set theory has been applied successfully to feature selection of discrete valued data [25,26,32]. Given a data set with discretized attribute values, it is possible to find a subset of the original attributes using rough set theory that are the most informative; all other attributes can be removed from the data set with minimal information loss. From the dimensionality reduction perspective, informative features are those that are most useful in determining classifications from their values [33,34].

One of the popular rough set based feature selection algorithms is quick reduct algorithm [24,35] in which the dependency or quality of approximation of single attribute is first calculated with respect to the class labels or decision attribute. After selecting the best attribute, other attributes are added to it to produce better quality. Additions of attributes are stopped when the final subset of attributes has the same quality as that of maximum possible quality of the data set or the quality of the selected attributes remains same. Other notable algorithms include discernibility matrix based method [36,37], dynamic reducts [38], and so forth. However, all these approaches are computationally very costly. The variable precision rough set model [39–41], tolerance rough sets [42,43], and probabilistic rough sets [44–46] are the extensions of the original rough set based knowledge representation. Different heuristic approaches based on rough set theory are also developed for feature selection [47,48]. Combining rough sets and genetic algorithms, different algorithms have been proposed in [49–51] to discover optimal or close to optimal subset of features.

In this paper, a new feature selection method is proposed to select a set of genes from microarray gene expression data by maximizing both relevance and significance of the selected genes. It employs rough set theory to compute the relevance and significance of the genes. Hence, the only information required in the proposed feature selection method is in the form of equivalence partitions for each gene, which can be automatically derived from the given microarray data set. This avoids the need for domain experts to provide information on the data involved and ties in with the advantage of rough sets is that it requires no information other than the data set itself. The use of both relevance and significance criteria for selecting genes with high predictive accuracy is theoretically justified based on the rough set theory. The importance of rough sets over mutual information is also established. The performance of the proposed approach is compared with that of existing approaches using the predictive accuracy of K-nearest neighbor rule and support vector machine on different microarray data sets.

The structure of the rest of this paper is as follows: Section 2 introduces the necessary notions of rough sets. The theoretical analysis on the relationships of dependency, relevance, and significance is presented in Section 3 using rough set theory. The proposed feature selection method is described in Section 4 for selecting relevant and significant genes from microarray data sets. Section 5 presents a methodology to compute rough set based relevance and significance criteria for continuous valued gene expression data set. A few case studies and a comparison with other related methods are presented in Section 6. Concluding remarks are given in Section 7.

2. Rough sets

The theory of rough sets begins with the notion of an approximation space, which is a pair $\langle \mathbb{U}, \mathbb{A} \rangle$, where $\mathbb{U}$ be a non-empty set, the universe of discourse, $\mathbb{U} = \{x_1, \ldots, x_i, \ldots, x_n\}$ and $\mathbb{A}$ is a family of attributes, also called knowledge in the

universe. V is the value domain of $\mathbb{A}$ and f is an information function $f : \mathbb{U} \times \mathbb{A} \to V$. An approximation space is also called an information system [22]. Any subset $\mathbb{P}$ of knowledge $\mathbb{A}$ defines an equivalence, also called indiscernibility, relation $IND(\mathbb{P})$ on $\mathbb{U}$

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

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

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

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

Given an arbitrary set $X \subseteq \mathbb{U}$, in general it may not be possible to describe X precisely in $\langle \mathbb{U}, \mathbb{A} \rangle$. One may characterize X by a pair of lower and upper approximations defined as follows [22]:

$$\underline{\mathbb{P}}(X) = \bigcup \{[x_i]_{\mathbb{P}} | [x_i]_{\mathbb{P}} \subseteq X\} \quad \text{and} \quad \overline{\mathbb{P}}(X) = \bigcup \{[x_i]_{\mathbb{P}} | [x_i]_{\mathbb{P}} \cap X \neq \emptyset\}. \tag{2}$$

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 $\langle \mathbb{U}, \mathbb{A} \rangle$ or simply called the rough set of X. The lower (respectively, upper) approximation $\underline{\mathbb{P}}(X)$ (respectively, $\overline{\mathbb{P}}(X)$) is interpreted as the collection of those elements of $\mathbb{U}$ that definitely (respectively, possibly) belong to X. The lower approximation is also called positive region sometimes, denoted as $POS_{\mathbb{P}}(X)$. A set X is said to be definable or exact in $\langle \mathbb{U}, \mathbb{A} \rangle$ iff $\underline{\mathbb{P}}(X) = \overline{\mathbb{P}}(X)$. Otherwise X is indefinable and termed as a rough set. $BN_{\mathbb{P}}(X) = \overline{\mathbb{P}}(X) \setminus \underline{\mathbb{P}}(X)$ is called a boundary set.

Definition 1. An information system $\langle \mathbb{U}, \mathbb{A} \rangle$ is called a decision table if the attribute set $\mathbb{A} = \mathbb{C} \cup \mathbb{D}$, where $\mathbb{C}$ is the condition attribute set and $\mathbb{D}$ is the decision attribute set. The dependency between $\mathbb{C}$ and $\mathbb{D}$ can be defined as [22]

$$\gamma_{\mathbb{C}}(\mathbb{D}) = \frac{|POS_{\mathbb{C}}(\mathbb{D})|}{|\mathbb{U}|}, \tag{3}$$

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

An important issue in data analysis is discovering dependency between attributes. Intuitively, a set of attributes $\mathbb{D}$ depends totally on a set of attributes $\mathbb{C}$, denoted as $\mathbb{C} \Rightarrow \mathbb{D}$, if all attribute values from $\mathbb{D}$ are uniquely determined by values of attributes from $\mathbb{C}$. If there exists a functional dependency between values of $\mathbb{D}$ and $\mathbb{C}$, then $\mathbb{D}$ depends totally on $\mathbb{C}$. The dependency can be defined in the following way:

Definition 2. Given $\mathbb{C}, \mathbb{D} \subseteq \mathbb{A}$, it is said that $\mathbb{D}$ depends on $\mathbb{C}$ in a degree κ, denoted as $\mathbb{C} \Rightarrow_{\kappa} \mathbb{D}$, if

$$\kappa = \gamma_{\mathbb{C}}(\mathbb{D}) = \frac{|POS_{\mathbb{C}}(\mathbb{D})|}{|\mathbb{U}|}, \quad \text{where } 0 \leq \kappa \leq 1. \tag{4}$$

If $\kappa = 1$, $\mathbb{D}$ depends totally on $\mathbb{C}$, if $0 < \kappa < 1$, $\mathbb{D}$ depends partially (in a degree κ) on $\mathbb{C}$, and if $\kappa = 0$, then $\mathbb{D}$ does not depend on $\mathbb{C}$ [22].

To what extent an attribute is contributing to calculate the dependency on decision attribute can be calculated by the significance of that attribute. The change in dependency when an attribute is removed from the set of condition attributes, is a measure of the significance of the attribute. The higher the change in dependency, the more significant the attribute is. If the significance is 0, then the attribute is dispensable.

Definition 3. Given $\mathbb{C}$, $\mathbb{D}$ and an attribute $\mathcal{A} \in \mathbb{C}$, the significance of the attribute $\mathcal{A}$ is defined as [22]:

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

3. Relationships of Max-Dependency, Max-Relevance, and Max-Significance

This section establishes the relationships among Max-Dependency, Max-Relevance, and Max-Significance using the rough set theory.

3.1. Max-Dependency

Let $\mathbb{C} = \{\mathcal{A}_1, \ldots, \mathcal{A}_i, \ldots, \mathcal{A}_j, \ldots, \mathcal{A}_m\}$ denotes the set of m condition attributes or features of a given data set. In terms of rough sets, the task of attribute or feature selection is to find a feature subset $\mathbb{S} \subseteq \mathbb{C}$ with $d < m$ features $\{\mathcal{A}_i\}$, which jointly have the largest dependency on the target class or decision attribute set $\mathbb{D}$. This scheme, called Max-Dependency, has the following form:

$$\max \mathcal{D}(\mathbb{S}, \mathbb{D}), \quad \mathcal{D} = \gamma_{\{\mathcal{A}_i, i=1,\ldots,d\}}(\mathbb{D}), \tag{6}$$

where $\gamma_{\{\mathcal{A}_i, i=1,\ldots,d\}}(\mathbb{D})$ represents the dependency between the feature subset $\mathbb{S} = \{\mathcal{A}_i, i = 1, \ldots, d\}$ and target class label $\mathbb{D}$ and is given by (4).

Obviously, when d equals 1, the solution is the feature that maximizes $\gamma_{\mathcal{A}_j}(\mathbb{D})$; $(1 \leq j \leq m)$. When $d > 1$, a simple incremental search scheme is to add one feature at one time. This type of selection is called the first order incremental search. By definition of first order search, it is assumed that $\mathbb{S}_{d-1}$, that is, the set of $d-1$ features, has already been obtained. The task is to select the optimal dth feature $\mathcal{A}_d$ from the set $\{\mathbb{C} - \mathbb{S}_{d-1}\}$ that contributes to the largest increase of $\gamma_{\mathbb{S}}(\mathbb{D})$. The quick reduct algorithm of Chouchoulas and Shen [35] is based on the principle of Max-Dependency.

The dependency $\mathcal{D}$ in (6) is represented by the dependency of (4), that is, $\mathcal{D} = \gamma_{\mathbb{S}_d}(\mathbb{D})$, where $\mathbb{S}_d = \{\mathbb{S}_{d-1}, \mathcal{A}_d\}$. Hence, from the definition of dependency in rough sets, the first order incremental search algorithm optimizes the following condition to select dth feature from the set $\{\mathbb{C} - \mathbb{S}_{d-1}\}$:

$$\max_{\mathcal{A}_j \in \{\mathbb{C} - \mathbb{S}_{d-1}\}} \{\gamma_{\{\mathbb{S}_{d-1}, \mathcal{A}_j\}}(\mathbb{D})\}, \tag{7}$$

which is equivalent to optimize the following condition given the set of selected features $\mathbb{S}_{d-1}$:

$$\max_{\mathcal{A}_j \in \{\mathbb{C} - \mathbb{S}_{d-1}\}} \{\gamma_{\{\mathbb{S}_{d-1}, \mathcal{A}_j\}}(\mathbb{D}) - \gamma_{\mathbb{S}_{d-1}}(\mathbb{D})\} = \max_{\mathcal{A}_j \in \{\mathbb{C} - \mathbb{S}_{d-1}\}} \{\sigma_{\mathbb{S}_d}(\mathbb{D}, \mathcal{A}_j)\}. \tag{8}$$

Obviously, the Max-Dependency is equivalent to either maximizing the joint dependency between selected feature set and the target class label or maximizing the significance of the candidate feature with respect to the already-selected features.

Despite the theoretical value of Max-Dependency, it is often hard to generate the resultant equivalence classes due to two difficulties in the high-dimensional space: the number of samples is often insufficient and the generation of resultant equivalence classes is usually an ill-posed problem. Another drawback of Max-Dependency is the slow computational speed. These problems are most pronounced for real life applications. If each feature has c categorical or discrete states and n samples, then d features could have a maximum $\min\{c^d, n\}$ equivalence classes. When the number of equivalence classes increases very quickly and gets comparable to the number of samples n, the joint dependency of these features cannot be estimated correctly. Hence, although Max-Dependency feature selection might be useful to select a very small number of features when n is large, it is not appropriate for real life applications where the aim is to achieve high classification accuracy with a reasonably compact set of features.

3.2. Max-Relevance and Max-Significance

As Max-Dependency criterion is hard to implement, an alternative is to select features based on maximal relevance criterion (Max-Relevance). Max-Relevance is to search features satisfying (9), which approximates $\mathcal{D}(\mathbb{S}, \mathbb{D})$ in (6) with the mean value of all dependency values between individual feature $\mathcal{A}_i$ and target class label $\mathbb{D}$:

$$\max \mathcal{R}(\mathbb{S}, \mathbb{D}), \quad \mathcal{R} = \frac{1}{|\mathbb{S}|} \sum_{\mathcal{A}_i \in \mathbb{S}} \gamma_{\mathcal{A}_i}(\mathbb{D}). \tag{9}$$

It is likely that features selected according to Max-Relevance could have rich redundancy, that is, the dependency among these features could be large. When two features highly depend on each other, the respective class discriminative power would not change much if one of them were removed. Therefore, the following maximal significance (Max-Significance) condition can be added to select mutually exclusive features:

$$\max \mathcal{S}(\mathbb{S}, \mathbb{D}), \quad \mathcal{S} = \frac{1}{|\mathbb{S}|(|\mathbb{S}| - 1)} \sum_{\substack{\mathcal{A}_i \neq \mathcal{A}_j \in \mathbb{S} \\ j > i}} \{\sigma_{\{\mathcal{A}_i, \mathcal{A}_j\}}(\mathbb{D}, \mathcal{A}_i) + \sigma_{\{\mathcal{A}_i, \mathcal{A}_j\}}(\mathbb{D}, \mathcal{A}_j)\}. \tag{10}$$

The criterion combining the above two constraints is called "maximal-relevance-maximal-significance" (MRMS). The operator $\Phi(\mathcal{R}, \mathcal{S})$ is defined to combine $\mathcal{R}$ and $\mathcal{S}$, and the following simplest form is considered to optimize $\mathcal{R}$ and $\mathcal{S}$ simultaneously:

$$\max \Phi(\mathcal{R}, \mathcal{S}), \quad \Phi = \mathcal{R} + \mathcal{S}. \tag{11}$$

In practice [8,10,16], incremental search methods can be used to find the near-optimal features defined by $\Phi(\cdot)$. Given the feature set $\mathbb{S}_{d-1}$ with $d-1$ features, the task is to select the dth feature from the set $\{\mathbb{C}-\mathbb{S}_{d-1}\}$. This is done by selecting the feature that maximizes $\Phi(\cdot)$. The respective incremental algorithm optimizes the following condition:

$$\max_{\mathcal{A}_j\in\{\mathbb{C}-\mathbb{S}_{d-1}\}}\left[\gamma_{\mathcal{A}_j}(\mathbb{D})+\frac{1}{d-1}\sum_{\mathcal{A}_i\in\mathbb{S}_{d-1}}\sigma_{\{\mathcal{A}_i,\mathcal{A}_j\}}(\mathbb{D},\mathcal{A}_j)\right]. \tag{12}$$

Hence, the combination of Max-Relevance and Max-Significance, that is, the MRMS criterion, is equivalent to maximizing the dependency between the candidate feature $\mathcal{A}_d$ and class label $\mathbb{D}$ as well as maximizing the average value of all significance values of the candidate feature $\mathcal{A}_d$ with respect to the already-selected feature $\mathcal{A}_i\in\mathbb{S}_{d-1}$.

The following conclusions can be drawn from the above discussions:

(i) Maximizing the first term of (12), that is, maximizing $\mathcal{R}(\mathbb{S},\mathbb{D})$ of (9), only leads to Max-Relevance. Clearly, the difference between Max-Relevance and Max-Dependency of (6) is rooted in the different definitions of dependency in terms of rough set theory. Eq. (9) does not consider the joint effect of features on the target class $\mathbb{D}$. On the contrary, Max-Dependency of (6) considers the dependency between the data distribution in multi-dimensional space and the target class $\mathbb{D}$. This difference is critical in many circumstances.
(ii) Maximizing the second term of (12) only, that is, maximizing $\mathcal{S}(\mathbb{S},\mathbb{D})$ of (10), is equivalent to searching mutually exclusive or independent features. This is not sufficient for selecting highly discriminative features.
(iii) The equivalence between Max-Dependency and Max-Significance indicates that Max-Significance is an optimal first order implementation of Max-Dependency.
(iv) Compared to Max-Dependency, the MRMS criterion avoids the estimation of resultant equivalence classes for multiple features. Instead, computing the resultant equivalence classes for two features could be much easier and more accurate. This also leads to a more efficient feature selection algorithm.

In this regard, it should be noted that the minimum-redundancy-maximum-relevance (mRMR) based feature selection algorithm [8,10] selects a subset of features from the whole feature set by maximizing the relevance and minimizing the redundancy of the selected features. However, the redundancy measure of the mRMR method does not take into account the supervised information of class labels, while both relevance and significance criteria of the proposed MRMS method are computed based on the class labels. Hence, the proposed MRMS method provides better performance than the existing mRMR method.

4. Proposed feature selection algorithm

In real data analysis such as microarray data, the data set may contain a number of insignificant features. The presence of such irrelevant and insignificant features may lead to a reduction in the useful information. Ideally, the selected features should have high relevance with the classes and high significance in the feature set. The features with high relevance are expected to be able to predict the classes of the samples. However, if insignificant features are present in the subset, they may reduce the prediction capability. A feature set with high relevance and high significance enhances the predictive capability. Accordingly, a measure is required that can enhance the effectiveness of feature set. In this paper, the rough set theory is used to select the relevant and significant features or genes from high dimensional microarray gene expression data sets.

4.1. Maximum Relevance-Maximum Significance

Let $\mathbb{C}=\{\mathcal{A}_1,\ldots,\mathcal{A}_i,\ldots,\mathcal{A}_j,\ldots,\mathcal{A}_m\}$ denotes the set of m features or genes of a given microarray data set and $\mathbb{S}$ is the set of selected genes. Define $\hat{f}(\mathcal{A}_i,\mathbb{D})$ as the relevance of the gene $\mathcal{A}_i$ with respect to the class labels $\mathbb{D}$ while $\tilde{f}(\mathcal{A}_i,\mathcal{A}_j)$ as the significance of the gene $\mathcal{A}_j$ with respect to the gene $\mathcal{A}_i$. The total relevance of all selected genes is, therefore, given by

$$\mathcal{J}_{\text{relev}}=\sum_{\mathcal{A}_i\in\mathbb{S}}\hat{f}(\mathcal{A}_i,\mathbb{D}), \tag{13}$$

while the total significance among the selected genes is

$$\mathcal{J}_{\text{signf}}=\sum_{\mathcal{A}_i\neq\mathcal{A}_j\in\mathbb{S}}\tilde{f}(\mathcal{A}_i,\mathcal{A}_j). \tag{14}$$

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

$$\mathcal{J} = \mathcal{J}_{\text{relev}} + \beta \mathcal{J}_{\text{signf}} = \sum_{\mathcal{A}_i \in \mathbb{S}} \hat{f}(\mathcal{A}_i, \mathbb{D}) + \beta \sum_{\substack{\mathcal{A}_i \neq \mathcal{A}_j \in \mathbb{S} \\ j > i}} \tilde{f}(\mathcal{A}_i, \mathcal{A}_j), \tag{15}$$

where β is a weight parameter. To solve the above problem, the following greedy algorithm is used.

(i) Initialize $\mathbb{C} \leftarrow \{\mathcal{A}_1, \ldots, \mathcal{A}_i, \ldots, \mathcal{A}_j, \ldots, \mathcal{A}_m\}, \mathbb{S} \leftarrow \emptyset$.
(ii) Calculate the relevance $\hat{f}(\mathcal{A}_i, \mathbb{D})$ of each feature or gene $\mathcal{A}_i \in \mathbb{C}$.
(iii) Select the gene $\mathcal{A}_i$ as the most relevant gene that has the highest relevance value $\hat{f}(\mathcal{A}_i, \mathbb{D})$. In effect, $\mathcal{A}_i \in \mathbb{S}$ and $\mathbb{C} = \mathbb{C} \setminus \mathcal{A}_i$.
(iv) Repeat the following two steps until the desired number of genes is selected.
(v) Calculate the significance of each of the remaining genes of $\mathbb{C}$ with respect to the selected genes of $\mathbb{S}$ and remove it from $\mathbb{C}$ if it has zero significance value with respect to any one of the selected genes.
(vi) From the remaining genes of $\mathbb{C}$, select gene $\mathcal{A}_j$ that maximizes the following condition:

$$\hat{f}(\mathcal{A}_j, \mathbb{D}) + \frac{\beta}{|\mathbb{S}|} \sum_{\mathcal{A}_i \in \mathbb{S}} \tilde{f}(\mathcal{A}_i, \mathcal{A}_j). \tag{16}$$

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

Both the relevance and significance of a gene are calculated based on the rough set theory. The relevance $\hat{f}(\mathcal{A}_i, \mathbb{D})$ of a gene $\mathcal{A}_i$ with respect to the class labels $\mathbb{D}$ is calculated using (4), while significance $\tilde{f}(\mathcal{A}_i, \mathcal{A}_j)$ of the gene $\mathcal{A}_j$ with respect to the already-selected gene $\mathcal{A}_i$ is computed using (5).

4.2. Computational complexity

The rough set based proposed gene selection method has low computational complexity with respect to the number of genes in the original microarray gene expression data set.

(i) The computation of the relevance of m genes is carried out in step 2 of the proposed algorithm, which has $\mathcal{O}(m)$ time complexity.
(ii) The selection of most relevant gene from the set of m genes, which is carried out in step 3, has also a complexity $\mathcal{O}(m)$.
(iii) There is only one loop in step 4 of the proposed gene selection method, which is executed $(d-1)$ times, where d represents the number of selected genes.
  (a) The computation of significance of a candidate gene with respect to the already-selected genes takes only a constant amount of time. If $\acute{m}$ represents the cardinality of the already-selected gene set, the total complexity to compute the significance of $(m - \acute{m})$ candidate genes, which is carried out in step 5, is $\mathcal{O}(m - \acute{m})$.
  (b) The selection of a gene from $(m - \acute{m})$ candidate genes by maximizing both relevance and significance, which is carried out in step 6, has also a complexity $\mathcal{O}(m - \acute{m})$.
  Hence, the total complexity to execute the loop $(d-1)$ times is $(\mathcal{O}((d-1)((m-\acute{m}) + (m-\acute{m})))) = \mathcal{O}(d(m-\acute{m}))$.

In effect, the selection of a set of d relevant and significant genes from the whole set of m genes using the proposed rough set based first order incremental search method has an overall computational complexity of $(\mathcal{O}(m)+\mathcal{O}(m)+\mathcal{O}(d(m-\acute{m}))) = \mathcal{O}(m)$ as $d, \acute{m} \ll m$.

5. Generation of equivalence classes

In microarray gene expression data, the class labels of samples are represented by discrete symbols, while the expression values of genes are continuous. Hence, to measure both relevance and significance of genes using rough set theory, the continuous expression values of a gene have to be divided into several discrete partitions to generate equivalence classes [16,52,53].

Different discretization methods such as discretization based on mean and standard deviation [16], equal frequency binning [52], Roughfication method [21], and so forth can be employed to discretize the continuous gene expression values. However, the inherent error that exists in discretization process is of major concern in the computation of relevance and significance of continuous valued genes [53]. To address this problem, a fuzzy set based discretization method is presented next to generate equivalence classes required to compute both relevance and significance of genes using rough set theory. In this context, it should be noted that the fuzzy-rough sets [26,54–57] and neighborhood rough sets [58] can handle continuous valued attributes without any discretization.

The family of normal fuzzy sets produced by a fuzzy partitioning of the universe of discourse can play the role of fuzzy equivalence classes. Given a finite set $\mathbb{U}$, $\mathbb{C}$ is a fuzzy condition attribute set in $\mathbb{U}$, which generates a fuzzy equivalence partition on $\mathbb{U}$. If c denotes the number of fuzzy equivalence classes generated by the fuzzy equivalence relation and n is the number of objects in $\mathbb{U}$, then c-partitions of $\mathbb{U}$ are sets of (cn) values $\{\mu_{ij}^{\mathbb{C}}\}$ that can be conveniently arrayed as a $(c \times n)$ matrix $\mathbb{M}_{\mathbb{C}} = [\mu_{ij}^{\mathbb{C}}]$, which is denoted by

$$\mathbb{M}_{\mathbb{C}} = \begin{pmatrix} \mu_{11}^{\mathbb{C}} & \mu_{12}^{\mathbb{C}} & \cdots & \mu_{1n}^{\mathbb{C}} \\ \mu_{21}^{\mathbb{C}} & \mu_{22}^{\mathbb{C}} & \cdots & \mu_{2n}^{\mathbb{C}} \\ \cdots & \cdots & \cdots & \cdots \\ \mu_{c1}^{\mathbb{C}} & \mu_{c2}^{\mathbb{C}} & \cdots & \mu_{cn}^{\mathbb{C}} \end{pmatrix} \tag{17}$$

subject to $\sum_{i=1}^{c} \mu_{ij}^{\mathbb{C}} = 1,\ \forall j$, and for any value of i, if $k = \arg\max_j\{\mu_{ij}^{\mathbb{C}}\}$, then $\max_j\{\mu_{ij}^{\mathbb{C}}\} = \max_l\{\mu_{lk}^{\mathbb{C}}\} > 0$, where $\mu_{ij}^{\mathbb{C}} \in [0, 1]$ represents the membership of object x_j in the ith fuzzy equivalence partition or class F_i. The above axioms should hold for every fuzzy equivalence partition, which correspond to the requirement that an equivalence class is nonempty. Obviously, this definition degenerates to the normal definition of equivalence classes when the equivalence relation is nonfuzzy.

Each row of the matrix $\mathbb{M}_{\mathbb{C}}$ is a fuzzy equivalence partition or class [32,59,60]. In the proposed gene selection method, the π function in one dimensional form is used to assign membership values to different fuzzy equivalence classes for the input genes. A fuzzy set with membership function $\pi(x; \bar{c}, \sigma)$ represents a set of points clustered around $\bar{c}$, where

$$\pi(x; \bar{c}, \sigma) = \begin{cases} 2\left(1 - \frac{\|x-\bar{c}\|}{\sigma}\right)^2 & \text{for } \frac{\sigma}{2} \le \|x - \bar{c}\| \le \sigma, \\ 1 - 2\left(\frac{\|x-\bar{c}\|}{\sigma}\right)^2 & \text{for } 0 \le \|x - \bar{c}\| \le \frac{\sigma}{2}, \\ 0 & \text{otherwise}, \end{cases} \tag{18}$$

where $\sigma > 0$ is the radius of the π function with $\bar{c}$ as the central point and $\|\cdot\|$ denotes the Euclidean norm. When the pattern x lies at the central point $\bar{c}$ of a class, then $\|x - \bar{c}\| = 0$ and its membership value is maximum, that is, $\pi(\bar{c}; \bar{c}, \sigma) = 1$. The membership value of a point decreases as its distance from the central point $\bar{c}$, that is, $\|x-\bar{c}\|$ increases. When $\|x-\bar{c}\| = (\frac{\sigma}{2})$, the membership value of x is 0.5 and this is called a crossover point [61]. The $(c \times n)$ matrix $\mathbb{M}_{\mathcal{A}_i}$, corresponding to the ith gene $\mathcal{A}_i$, can be calculated from the c-fuzzy equivalence classes of the objects $x = \{x_1, \ldots, x_j, \ldots, x_n\}$, where

$$\mu_{kj}^{\mathcal{A}_i} = \frac{\pi(x_j; \bar{c}_k, \sigma_k)}{\sum_{l=1}^{c} \pi(x_j; \bar{c}_l, \sigma_l)}. \tag{19}$$

In effect, each position $\mu_{kj}^{\mathcal{A}_i}$ of the matrix $\mathbb{M}_{\mathcal{A}_i}$ must satisfy the following conditions:

$$\mu_{kj}^{\mathcal{A}_i} \in [0, 1]; \quad \sum_{k=1}^{c} \mu_{kj}^{\mathcal{A}_i} = 1, \quad \forall j \text{ and for any value of } k, \text{ if}$$
$$s = \arg\max_j\{\mu_{kj}^{\mathcal{A}_i}\}, \text{ then } \max_j\{\mu_{kj}^{\mathcal{A}_i}\} = \max_l\{\mu_{ls}^{\mathcal{A}_i}\} > 0.$$

After the generation of the matrix $\mathbb{M}_{\mathcal{A}_i}$ corresponding to the gene $\mathcal{A}_i$, the object x_j is assigned to one of the c equivalence classes based on the maximum value of memberships of the object in different equivalence classes that follows next:

$$x_j \in F_p, \quad \text{where } p = \arg\max_k\{\mu_{kj}^{\mathcal{A}_i}\}.$$

Each input real valued gene in quantitative form can be assigned to different fuzzy equivalence classes in terms of membership values using the π fuzzy set with appropriate $\bar{c}$ and σ. The centers and radii of the π functions along each gene axis are determined automatically from the distribution of the training patterns. In the proposed gene selection algorithm, three fuzzy equivalence classes ($c = 3$), namely, low, medium, and high are considered. These three equivalence classes correspond to under-expression, base-line, and over-expression of continuous valued genes, respectively. Corresponding to three fuzzy sets low, medium, and high, the following relations hold:

$$\bar{c}_1 = \bar{c}_{\text{low}}(\mathcal{A}_i); \quad \bar{c}_2 = \bar{c}_{\text{medium}}(\mathcal{A}_i); \quad \bar{c}_3 = \bar{c}_{\text{high}}(\mathcal{A}_i); \quad \sigma_1 = \sigma_{\text{low}}(\mathcal{A}_i); \quad \sigma_2 = \sigma_{\text{medium}}(\mathcal{A}_i); \quad \sigma_3 = \sigma_{\text{high}}(\mathcal{A}_i).$$

The parameters $\bar{c}$ and σ of each π fuzzy set are computed according to the following procedure [61]. Let $\bar{m}_i$ be the mean of the objects $x = \{x_1, \ldots, x_j, \ldots, x_n\}$ along the ith gene $\mathcal{A}_i$. Then $\bar{m}_{i_l}$ and $\bar{m}_{i_h}$ are defined as the mean along the ith gene of the objects having co-ordinate values in the range $[\mathcal{A}_{i_{\min}}, \bar{m}_i)$ and $(\bar{m}_i, \mathcal{A}_{i_{\max}}]$, respectively, where $\mathcal{A}_{i_{\max}}$ and $\mathcal{A}_{i_{\min}}$ denote the upper and lower bounds of the dynamic range of gene $\mathcal{A}_i$ for the training set. For three fuzzy sets low, medium, and high, the centers and corresponding radii are computed as follows:

$$\bar{c}_{\text{low}}(\mathcal{A}_i) = \bar{m}_{i_l}; \quad \bar{c}_{\text{medium}}(\mathcal{A}_i) = \bar{m}_i; \quad \bar{c}_{\text{high}}(\mathcal{A}_i) = \bar{m}_{i_h};$$

$$\sigma_{\text{low}}(\mathcal{A}_i) = 2(\bar{c}_{\text{medium}}(\mathcal{A}_i) - \bar{c}_{\text{low}}(\mathcal{A}_i)); \quad \sigma_{\text{high}}(\mathcal{A}_i) = 2(\bar{c}_{\text{high}}(\mathcal{A}_i) - \bar{c}_{\text{medium}}(\mathcal{A}_i)); \quad \sigma_{\text{medium}}(\mathcal{A}_i) = \eta \times \frac{A}{B};$$

$$\text{where} \quad A = \{\sigma_{\text{low}}(\mathcal{A}_i)(\mathcal{A}_{i_{\max}} - c_{\text{medium}}(\mathcal{A}_i)) + \sigma_{\text{high}}(\mathcal{A}_i)(c_{\text{medium}}(\mathcal{A}_i) - \mathcal{A}_{i_{\min}})\}; \quad B = \{\mathcal{A}_{i_{\max}} - \mathcal{A}_{i_{\min}}\},$$

where η is a multiplicative parameter controlling the extent of the overlapping. The distribution of the patterns or objects along each gene axis is taken into account, while computing the corresponding centers and radii of the fuzzy sets. Also, the amount of overlap between the three fuzzy sets can be different along the different axis, depending on the distribution of the objects or patterns.

6. Experimental results

The performance of the proposed rough set based maximum relevance-maximum significance (MRMS) method is extensively studied and compared with that of some existing algorithms, namely, minimum redundancy-maximum relevance (mRMR) framework [8], Quick Reduct algorithm [35], Discernibility Matrix based approach [37], Roughfication [21], the methods proposed by Valdes and Barton [30] and Fang and Busse [29]. The performance of the MRMS method is also compared with that of Max-Dependency and Max-Relevance criteria, along with the comparison between fuzzy and crisp equivalence classes [16,52]. The proposed MRMS algorithm is implemented in C language and run in LINUX environment having machine configuration Pentium IV, 2.8 GHz, 1 MB cache, and 1 GB RAM.

To analyze the performance of different algorithms, the experimentation is done on five cancer and two arthritis microarray data sets. For each data set, 50 top-ranked genes is selected for analysis, and each data set is pre-processed by standarizing each sample to zero mean and unit variance. The major metrics for evaluating the performance of different algorithms are the classification accuracy of K-nearest neighbor (K-NN) rule and support vector machine (SVM). To compute the prediction accuracy of both SVM and K-NN rule, both leave-one-out cross-validation (LOOCV) and 10-fold cross-validation (10-fold CV) are performed on each gene expression data set.

6.1. Gene expression data sets

In this paper, publicly available five cancer and two arthritis data sets are used. Since binary classification is a typical and fundamental issue in diagnostic and prognostic prediction of both cancer and arthritis, different methods are compared using the following binary class data sets.

(i) *Breast Cancer:* The breast cancer data set contains expression levels of 7129 genes in 49 breast tumor samples [62]. The samples are classified according to their estrogen receptor (ER) status: 25 samples are ER positive while other 24 samples are ER negative.
(ii) *Leukemia:* It is an affymetrix high density oligonucleotide array that contains 7070 genes and 72 samples from two classes of leukemia [1]: 47 acute lymphoblastic leukemia and 25 acute myeloid leukemia.
(iii) *Colon Cancer:* The colon cancer data set contains expression levels of 2000 genes and 62 samples from two classes [63]: 40 tumor and 22 normal colon tissues.
(iv) *Lung Cancer:* This data set contains 181 tissue samples: among them 31 are malignant pleural mesothelioma and rest 150 adenocarcinoma of the lung [64]. Each sample is described by the expression levels of 12,533 genes.
(v) *Prostate Cancer:* In this data set, 136 samples are grouped into two classes: 77 prostate tumor and 59 prostate normal samples [65]. Each sample contains 12,600 genes.
(vi) *Rheumatoid Arthritis versus Osteoarthritis (RAOA):* The RAOA data set consists of gene expression profiles of thirty patients: 21 with RA and 9 with OA [66]. The Cy5-labeled experimental cDNA and the Cy3 labeled common reference sample were pooled and hybridized to the lymphochips containing $\sim$18,000 cDNA spots representing genes of relevance in immunology [66].
(vii) *Rheumatoid Arthritis versus Healthy Controls (RAHC):* The RAHC data set consists of gene expression profiling of peripheral blood cells from 32 patients with RA, three patients with probable RA and 15 age and sex matched healthy controls performed on microarrays with a complexity of $\sim$26K unique genes (43K elements) [67].

6.2. Class prediction methods

Following two quantitative indices are used to evaluate the performance of different methods with respect to seven microarray data sets.

6.2.1. Support vector machine

The support vector machine (SVM) [68] 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

Table 1
Comparative performance of rough sets and mutual information using LOOCV.

Microarray	Quantitative	Rough sets + D1		Mutual information + D1		Rough sets + D2		Mutual information + D2	
data set	measures	Accuracy	Gene	Accuracy	Gene	Accuracy	Gene	Accuracy	Gene
Breast	SVM	100	18	97.96	11	100	8	97.9	10
	K-NN	100	45	93.88	6	98.0	27	95.9	6
Leukemia	SVM	97.2	22	98.61	19	100	14	98.6	24
	K-NN	98.6	47	95.83	25	100	37	98.6	42
Colon	SVM	87.1	5	87.1	5	85.5	33	80.6	15
	K-NN	83.9	3	88.71	40	77.4	23	79.0	46
Lung	SVM	100	34	99.45	2	100	8	99.5	25
	K-NN	100	38	99.45	2	99.5	9	99.5	20
Prostate	SVM	89.7	44	96.32	47	94.9	48	94.9	12
	K-NN	88.2	7	92.65	27	94.9	23	94.9	18
RAOA	SVM	100	5	100	7	100	8	100	6
	K-NN	100	3	100	11	100	12	96.7	7
RAHC	SVM	90	20	98	10	100	33	96	13
	K-NN	100	11	100	16	100	21	98	49

good generalization properties. A key factor in the SVM is to use kernels to construct nonlinear decision boundary. In the present work, linear kernels are used.

6.2.2. K-nearest neighbor rule

The K-nearest neighbor (K-NN) rule [69] is used for evaluating the effectiveness of the reduced gene set for classification. It classifies samples based on closest training samples in the feature space. A sample is classified by a majority vote of its K-neighbors, with the sample being assigned to the class most common amongst its K-nearest neighbors. The value of K, chosen for the K-NN, is the square root of the number of samples in training set.

6.3. Importance of rough sets

In the proposed MRMS method, both the relevance and significance of a gene are calculated based on the rough set theory. The relevance of a gene with respect to the class labels is calculated using (4), while significance of a gene with respect to the already-selected gene is computed using (5). However, other measures such as mutual information can also be used to compute both relevance and significance of a gene. In order to establish the importance of rough sets over mutual information, extensive experimental results are reported in Table 1 for seven microarray data sets. Subsequent discussions analyze the results with respect to the classification accuracy of both SVM and K-NN rule. The value of β is set to 1.0 for the MRMS criterion and the equivalence classes are generated by two discretization methods: using mean-standard deviation (D1) [16] and equal frequency binning (D2) [52].

From the results reported in Table 1, it is seen that the performance of rough sets is better than that of mutual information in most of the cases. Out of total 28 cases, the MRMS criterion achieves significantly better results for rough sets in 19 cases. However, the mutual information provides better accuracy of the SVM for leukemia, prostate cancer, and RAHC data sets and that of the K-NN for colon and prostate cancer data sets using the method D1. On the other hand, the rough set based approach provides same accuracy of the SVM and K-NN with higher number of genes for prostate cancer data set, same accuracy of the SVM with higher number of genes for RAOA data set, and lower accuracy of the K-NN for colon cancer data set using the method D2.

6.4. Effectiveness of MRMS criterion

To establish the effectiveness of the proposed MRMS criterion based gene selection method over Max-Dependency and Max-Relevance criteria, extensive experimental results are reported in Table 2 for seven microarray data sets. Subsequent discussions analyze the results with respect to the classification accuracy of both SVM and K-NN rule. The best results obtained using Max-Dependency and Max-Relevance criteria on these data sets are also presented in this table for the sake of comparison. The value of β varies from 0.0 to 1.0 for the MRMS criterion and the equivalence classes are generated by two discretization methods: using mean-standard deviation (D1) [16] and equal frequency binning (D2) [52]. In this context, it should be noted that the Max-Relevance criterion is equivalent to the proposed MRMS criterion with $\beta = 0.0$, while the quick reduct algorithm of Chouchoulas and Shen [35] follows the Max-Dependency criterion.

6.4.1. Optimum value of β

The parameter β regulates the relative importance of the significance of the candidate gene with respect to the already-selected genes and the relevance with the output class. If β is zero, only the relevance with the output class is considered for each gene selection. If β increases, this measure is incremented by a quantity proportional to the total significance with respect to the already-selected genes. The presence of a β value larger than zero is crucial in order to obtain good results. If

Table 2
Comparative performance of Max-Dependency, Max-Relevance, and proposed algorithm using LOOCV.

Microarray	Quantitative	Discretization	Max-Dependency		Max-Relevance		MRMS ($\beta = 1.0$)		MRMS ($0.0 < \beta < 1.0$)		
data set	measures	procedure	Accuracy	Gene	Accuracy	Gene	Accuracy	Gene	Accuracy	Gene	Value of β
Breast	SVM	Method: D1	85.7	3	98.0	11	100	18	100	18	0.6–0.9
		Method: D2	87.8	3	100	9	100	8	100	8	0.1–0.9
	K-NN	Method: D1	83.7	2	98.0	17	100	45	100	45	0.8–0.9
		Method: D2	91.8	3	91.8	8	98.0	27	98.0	27	0.1–0.9
Leukemia	SVM	Method: D1	100	3	97.2	32	97.2	22	98.6	36	0.1
		Method: D2	87.5	2	98.6	43	100	14	100	14	0.1–0.8
	K-NN	Method: D1	98.6	2	98.6	43	98.6	47	100	50	0.1–0.3
		Method: D2	90.3	3	97.2	25	100	37	100	37	0.1–0.9
Colon	SVM	Method: D1	80.7	2	80.7	23	87.1	5	87.1	5	0.9
		Method: D2	62.9	1	74.2	4	85.5	33	85.5	33	0.1–0.9
	K-NN	Method: D1	80.7	3	82.3	50	83.9	3	85.5	9	0.9
		Method: D2	64.5	1	69.4	7	77.4	23	77.4	23	0.1–0.9
Lung	SVM	Method: D1	99.5	3	99.5	7	100	34	100	34	0.6–0.9
		Method: D2	98.3	3	99.5	31	100	8	100	8	0.1–0.9
	K-NN	Method: D1	99.5	3	99.5	42	100	38	100	39	0.9
		Method: D2	97.8	3	99.5	22	99.5	9	99.5	9	0.1–0.9
Prostate	SVM	Method: D1	84.6	4	81.6	47	89.7	44	89.7	44	0.9
		Method: D2	56.6	1	62.5	6	94.9	48	94.9	48	0.1–0.9
	K-NN	Method: D1	88.2	4	91.2	5	88.2	7	88.2	7	0.1–0.9
		Method: D2	55.8	1	63.9	25	94.9	23	94.9	23	0.1–0.9
RAOA	SVM	Method: D1	86.7	1	90.0	50	100	5	100	3	0.5–0.6
		Method: D2	73.3	2	96.7	16	100	8	100	4	0.2
	K-NN	Method: D1	90.0	2	90.0	2	100	3	100	3	0.7–0.9
		Method: D2	70.0	2	90.0	6	100	12	100	12	0.8–0.9
RAHC	SVM	Method: D1	70.0	1	94.0	16	90.0	20	94.0	36	0.1–0.4
		Method: D2	70.0	1	96.0	48	100	33	100	12	0.6
	K-NN	Method: D1	84.0	3	90.0	11	100	11	100	11	0.5–0.9
		Method: D2	82.0	3	86.0	11	100	21	100	12	0.8–0.9

the significance between genes is not taken into account, selecting the genes with the highest relevance with respect to the output class may tend to produce a set of redundant genes that may leave out useful complementary information.

The values of β for which the proposed MRMS criterion based gene selection algorithm achieves its best performance are reported in Table 2. From the results reported in this table, it is seen that the MRMS criterion attains its best performance at $\beta = 0.9$ for breast, colon, lung, and prostate cancer data sets using both SVM and K-NN rule, and for RAOA and RAHC data sets using only K-NN rule. On the other hand, the proposed algorithm provides its best results at $\beta = 0.1$ for leukemia data set using both SVM and K-NN rule and for RAHC data set using only the SVM. Hence, the MRMS criterion achieves its best performance for $0.1 \leq \beta \leq 0.9$ irrespective of the data sets and classifiers used.

6.4.2. Comparative performance analysis

From the results reported in Table 2, it is seen that the performance of proposed MRMS criterion is better than that of Max-Dependency and Max-Relevance criteria in most of the cases. Out of total 28 cases, the MRMS criterion achieves significantly better results than Max-Dependency or Max-Relevance in 25 cases. However, the Max-Dependency criterion provides better accuracy of the SVM for leukemia data set and same accuracy of the K-NN rule with lower number of genes for prostate cancer data set than the MRMS criterion. Also, the Max-Relevance criterion achieves better accuracy of the K-NN rule for prostate cancer data set and same accuracy of the SVM with lower number of genes for RAHC data set than the MRMS criterion. That is, both Max-Dependency and Max-Relevance criteria are useful to select a very small number of genes, but not appropriate to achieve high classification accuracy. Hence, the combination of Max-Relevance and Max-Significance, that is, the MRMS criterion, must be used to get a reduced set of genes with high classification accuracy.

6.5. Effectiveness of fuzzy equivalence classes

In order to improve the performance of proposed MRMS criterion based gene selection method, three π functions in one dimensional form are used to generate three equivalence classes, namely, low, medium, and high. The multiplicative parameter η controls the overlapping between three fuzzy equivalence classes low and medium or medium and high. Keeping the values of σ_{low} and σ_{high} fixed, the amount of overlapping among three π functions can be altered varying σ_{medium}. As η is decreased, the radius σ_{medium} decreases around $\bar{c}_{\text{medium}}$ such that ultimately there is insignificant overlapping between three π functions low and medium or medium and high. This implies that certain regions along the ith gene axis $\mathcal{A}_i$ go under-represented such that three membership values corresponding to three fuzzy sets low, medium, and high attain small values. Note that the particular choice of the values of σs and $\bar{c}$s ensure that for any pattern x_j along the ith gene axis $\mathcal{A}_i$, at least one of membership values should be greater than 0.5. On the other hand, as η is increased the radius σ_{medium} increases around $\bar{c}_{\text{medium}}$ such that the amount of overlapping between the three π functions increases.

Table 3
Best performance of proposed algorithm on seven data sets using LOOCV.

Microarray	SVM				*K*-NN			
data set	Value of β	Accuracy	Gene	Value of η	Value of β	Accuracy	Gene	Value of η
Breast	0.3–0.8	100	6	0.8	0.6–0.7	100	6	1.2
	0.0	100	7	0.8	0.0	100	10	1.4
Leukemia	0.1	100	4	1.5–1.6	0.1–0.2	100	3	1.7
	0.0	100	4	1.4–1.6	0.0	100	3	1.7
Colon	0.1–1.0	90.3	35	0.8	0.1–0.6	90.3	35	0.9
	0.0	88.7	21	0.6	0.0	88.7	20	0.6
Lung	0.7	100	9	1.3	1.0	100	4	0.9
	0.0	100	14	1.1	0.0	100	10	0.5–0.6
Prostate	0.9	96.3	43	1.3	0.9–1.0	95.6	6	2.0
	0.0	93.4	50	0.5	0.0	91.9	2	1.3
RAOA	0.6–1.0	100	4	0.8, 1.0	0.7–1.0	100	8	0.9
	0.0	100	30	0.5	0.0	93.3	2	0.7
RAHC	1.0	100	18	0.6	1.0	100	5	0.5
	0.0	100	28	0.6	0.0	98.0	22	0.6

To establish the effectiveness of fuzzy equivalence classes over the crisp equivalence classes and to find out the corresponding optimum values of both η and β, the extensive experimentation is carried out on seven microarray data sets. The value of β ranges from 0.0 to 1.0, while the value of η varies from 0.5 to 2.0.

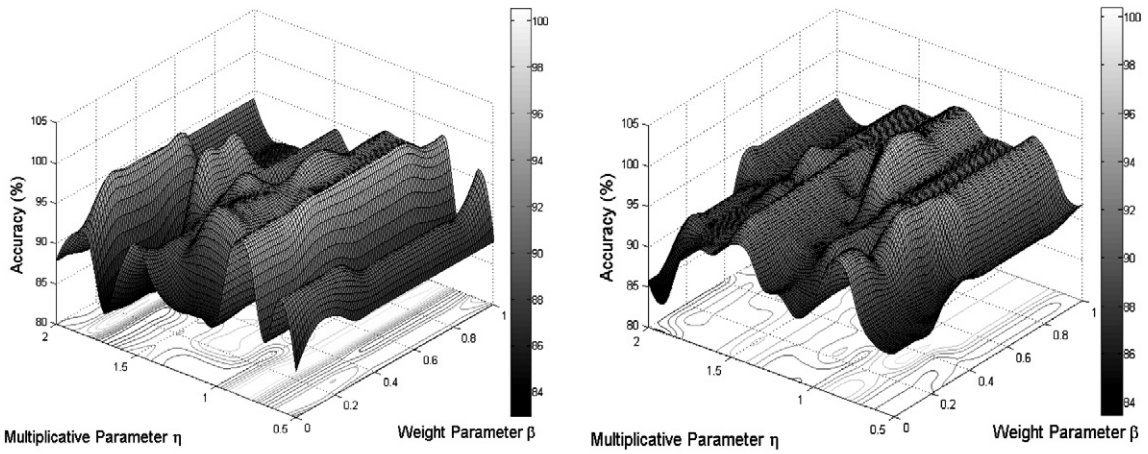
6.5.1. Variable number of selected genes

Table 3 presents the best performance of the proposed MRMS based gene selection algorithm for different data sets using fuzzy equivalence classes. The results and subsequent discussions are presented in this table with respect to the predictive accuracy of both SVM and *K*-NN rule. The values of β and η for which the best performance of the proposed algorithm is achieved are also reported in this table, along with the number of selected genes. From the results reported in Table 3, it is seen that the proposed algorithm with $\beta \neq 0.0$ provides better or comparable classification accuracy with lower number of selected genes than that of $\beta = 0.0$ in most of the cases. Only for leukemia, the performance of the proposed algorithm with $\beta = 0.1$ is same as that of $\beta = 0.0$. The corresponding values of η indicate that very large or very small amounts of overlapping among the three equivalence classes of input gene are found to be undesirable for $\beta > 0.0$.

6.5.2. Fixed number of selected genes

Figs. 1–7 present the performance of the proposed gene selection algorithm on five cancer and two arthritis microarray data sets for fixed number of genes. The results and subsequent discussions are presented in these figures for different values of β and η with respect to the predictive accuracy of both SVM and *K*-NN rule. For each data set, the number of selected genes is fixed through extensive experimentation in such a way that the classification accuracy of both SVM and *K*-NN rule attains its highest value.

From the results reported in Figs. 1–7, it is seen that as the value of β increases, the classification accuracy of both SVM and *K*-NN rule increases. On the other hand, the performance decreases for very high or very low values of η. The proposed rough



(a) SVM with $d = 6$ (b) K-NN with $d = 6$

Fig. 1. Variation of classification accuracy with respect to multiplicative parameter η and weight parameter β for breast cancer.

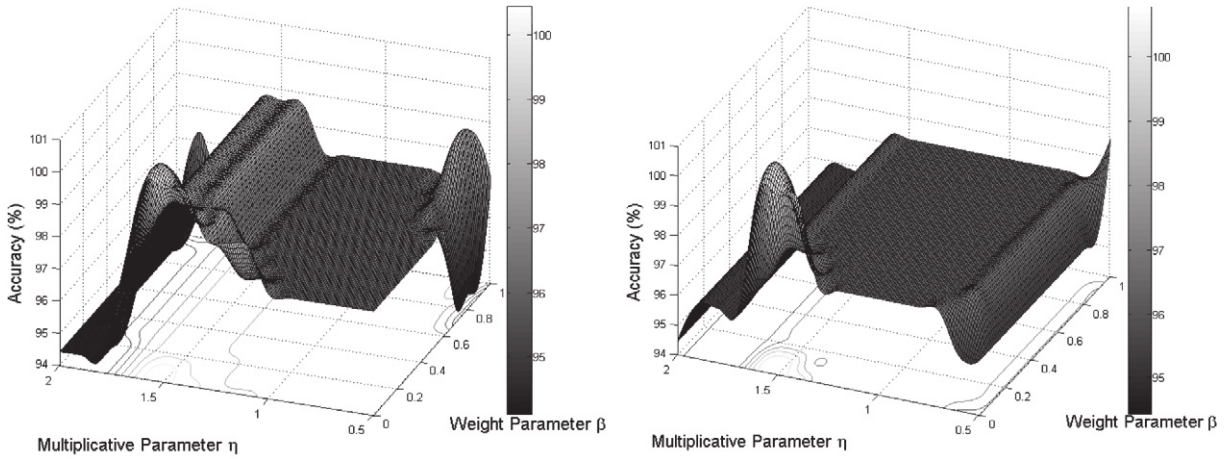


Fig. 2. Variation of classification accuracy with respect to multiplicative parameter η and weight parameter β for leukemia.

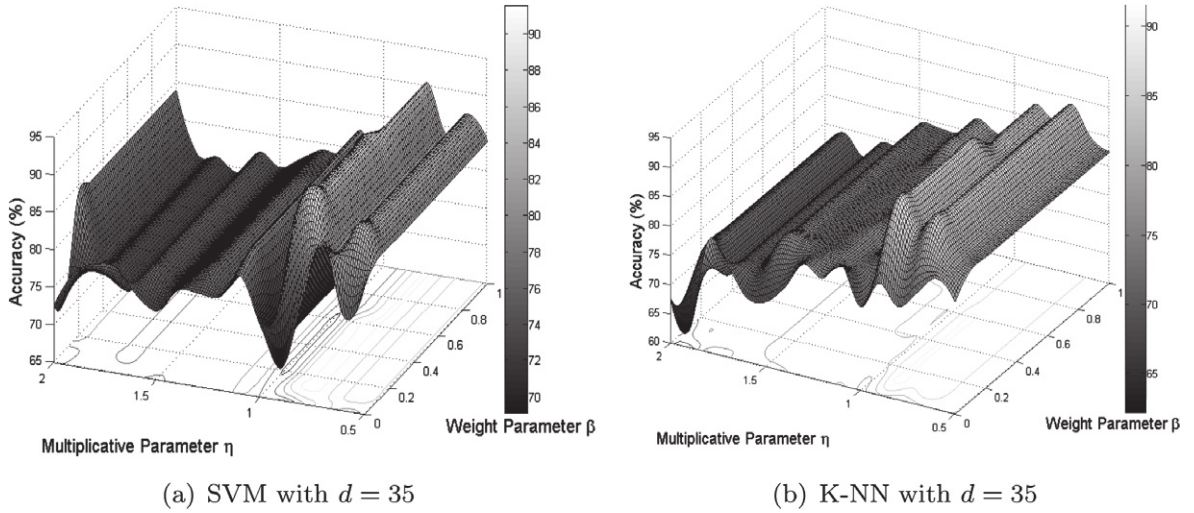


Fig. 3. Variation of classification accuracy with respect to multiplicative parameter η and weight parameter β for colon cancer.

set based gene selection algorithm achieves its best performance for $\beta > 0.0$ with respect to the classification accuracy of both SVM and K-NN rule. The MRMS criterion achieves 100% accuracy for leukemia and 90.3% accuracy for colon at $\beta = 0.1$, 100% accuracy for breast and 90.3% accuracy for colon at $\beta = 0.6$, 100% accuracy for breast, lung, and RAOA data at $\beta = 0.7$, and 100% accuracy for RAOA and RAHC data at $\beta = 0.9$, irrespective of the classifiers used. For prostate cancer data, it attains 96.3% and 95.6% accuracy at $\beta = 0.9$ using the SVM and K-NN rule, respectively. All these results are obtained for $0.7 \leq \eta \leq 1.7$. In other words, the best performance of proposed method is achieved when the relevance of each gene is incremented by at least 10% of the total significance with respect to the already-selected genes. However, the performance of the proposed method at $\beta = 0.0$ is same as that of $\beta = 0.1$ for leukemia data set using both SVM and K-NN rule. The important results corresponding to Figs. 1–7 are also summarized in Table 4.

From the results reported in Tables 3 and 4 and Figs. 1–7, it is seen that, for a particular number of selected genes, the predictive accuracy of both SVM and K-NN rule for $\beta > 0.0$ is higher compared to that of $\beta = 0.0$, irrespective of the microarray gene expression data sets used. Moreover, it is seen that very large or very small amounts of overlapping among the three π fuzzy equivalence classes of the input genes lead to undesirable results for $\beta > 0.0$.

6.5.3. Performance of fuzzy equivalence classes

Finally, Table 5 reports the comparative performance of crisp and fuzzy equivalence classes with respect to the classification accuracy of both SVM and K-NN rule. The crisp equivalence classes are generated by two discretization methods: using mean-standard deviation (D1) [16] and equal frequency binning (D2) [52]. From the results reported in Table 5, it is seen

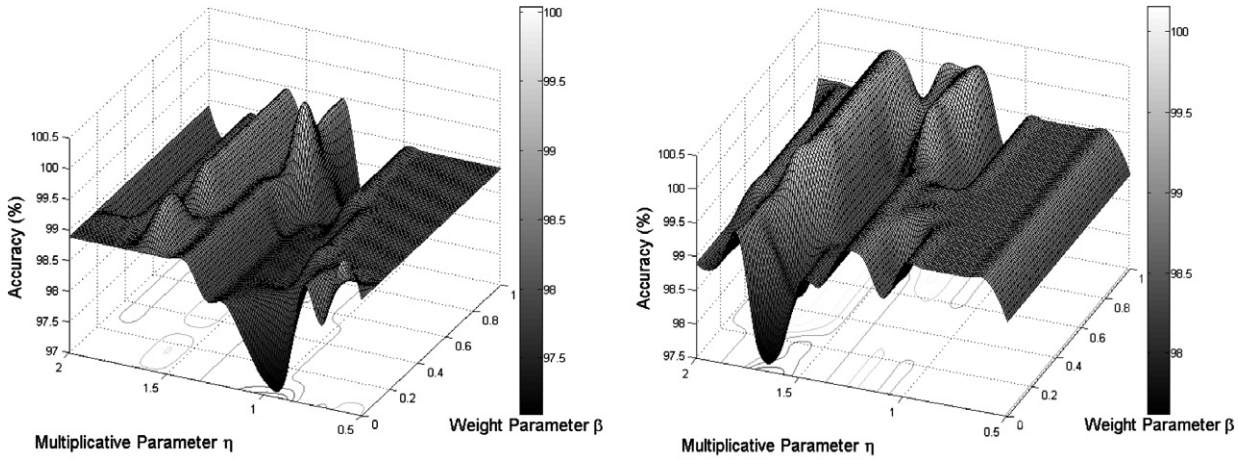


Fig. 4. Variation of classification accuracy with respect to multiplicative parameter η and weight parameter β for lung cancer.

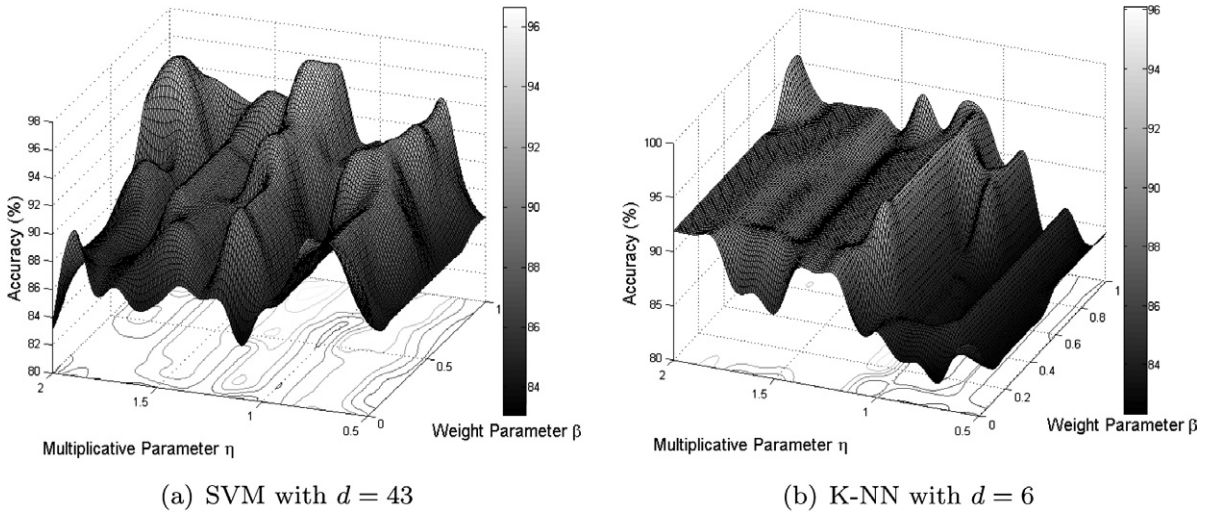


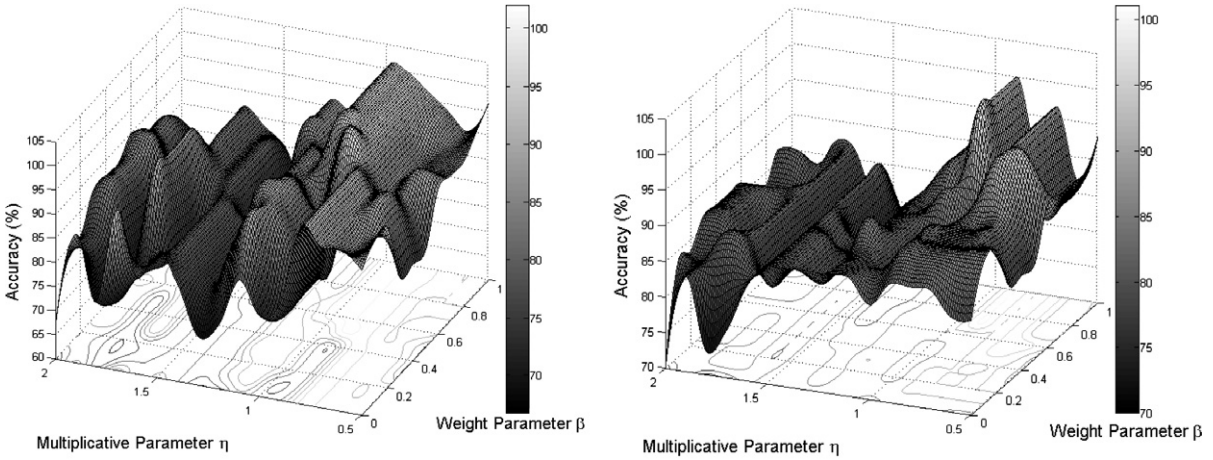
Fig. 5. Variation of classification accuracy with respect to multiplicative parameter η and weight parameter β for prostate.

that the proposed gene selection algorithm with fuzzy equivalence classes performs better than that with crisp equivalence classes in most of the cases. However, only for RAOA data set, the proposed algorithm with crisp equivalence classes produced by the method D1 attains same accuracy as that with fuzzy equivalence classes with lower number of genes. On the other hand, the discretization method D2 achieves same accuracy as that of fuzzy equivalence classes with lower number of genes for lung and RAHC data sets using the SVM.

6.6. Comparative performance analysis of different algorithms

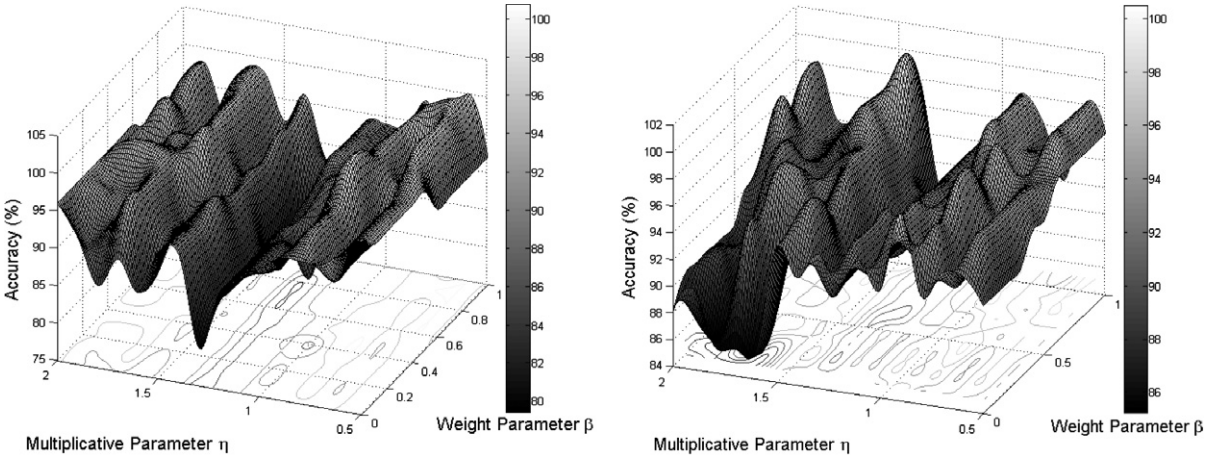
Finally, the best results of different algorithms on seven microarray data sets are presented in Tables 7–9, while Table 6 reports the results considering the whole gene set. Subsequent discussions analyze the results with respect to the prediction accuracy of the SVM and K-NN rule. The best performance of some existing algorithms such as mRMR [8], Quick Reduct algorithm [35], Discernibility Matrix based approach [37], Roughfication [21], the methods proposed by Valdes and Barton [30] and Fang and Busse [29], is provided on same data sets for the sake of comparison.

Both LOOCV and 10-fold CV are performed on each data set. In case of 10-fold CV, the means and standard deviations of the classification accuracy of the SVM and K-NN rule are computed for all data sets. Tests of significance are performed for the inequality of means (of the classification accuracy of both SVM and K-NN rule) obtained using the proposed MRMS method and the other related algorithms compared. Since both mean pairs and the variance pairs are unknown and different, a generalized version of t-test is used here. The above problem is the classical Behrens–Fisher problem in hypothesis testing.



(a) SVM with $d = 8$ (b) K-NN with $d = 8$

Fig. 6. Variation of classification accuracy with respect to multiplicative parameter η and weight parameter β for RAOA data.



(a) SVM with $d = 36$ (b) K-NN with $d = 36$

Fig. 7. Variation of classification accuracy with respect to multiplicative parameter η and weight parameter β for RAHC data.

The test statistic, which is described and tabled in [70], is of the form

$$t = \frac{\mu_1 - \mu_2}{\sqrt{\lambda_1 {\sigma_1}^2 + \lambda_2 {\sigma_2}^2}}, \tag{20}$$

where μ_1, μ_2 are the means, σ_1, σ_2 the standard deviations, and $\lambda_1 = 1/n_1$, $\lambda_2 = 1/n_2$, n_1, n_2 are the number of observations. Tables 7 and 9 report the individual means and standard deviations, and the value of test statistic computed. The corresponding tabled value is 1.81 at an error probability level of 0.05. If the computed value is greater than the tabled value, the means are significantly different.

6.6.1. Results on full gene set

The classification accuracy of both SVM and K-NN rule is reported in Table 6 considering the whole gene set. That is, the K-NN rule and SVM are used to classify the samples of each microarray data set considering all genes of the data set and the performance is compared with that of different feature selection algorithms, which are reported in Tables 7–9. The results reported in Table 6 indicate that if all genes are considered for sample classification, the samples from different classes may not be well separated with respect to the K-NN rule and SVM. However, from the results reported in Tables 7–9, it can be seen that when a gene or feature selection algorithm selects a set of genes from the whole gene set considering the relevance, redundancy, or significance criteria, the genes those have high relevance with respect to the class labels are only selected. In effect, the samples from different classes with reduced gene set become well separated, which leads to higher classification

Table 4
Optimum values of β and η for different data sets using LOOCV.

Microarray	SVM		K-NN	
data set	Accuracy	Values of (β, η)	Accuracy	Values of (β, η)
Breast	100	({0.3–0.8}, 0.8)	100	({0.6–0.7}, 1.2)
$d = 6$	98.0	(0.0, {0.8–0.9})	95.9	(0.0, {0.8–0.9})
Leukemia	100	(0.1, {1.5–1.6})	100	(0.1, {1.5–1.6})
$d = 4$	100	(0.0, {1.4–1.6})	100	(0.0, {1.4–1.6})
Colon	90.3	({0.1–1.0}, 0.8)	90.3	({0.1–0.6}, 0.9)
$d = 35$	87.1	(0.0, 0.7)	82.3	(0.0, 0.6)
Lung	100	(0.7, 1.3)	100	(0.3, 1.7), ({0.4–0.6}, {1.6–1.7}), ({0.7–0.8}, 1.3), ({0.7–0.8}, {1.6–1.7}), (0.9, {1.2–1.3}), (0.9, {1.6–1.7}), (1.0, {1.2–1.4}), (1.0, {1.6–1.7})
$d = 9$	99.5	(0.0, 0.6)	99.5	(0.0, {0.6–1.0}), (0.0, {1.2–1.3}), (0.0, 1.8)
Prostate	96.3	(0.9, 1.3)	95.6	({0.9–1.0}, 2.0)
$d = 43/6$	91.9	(0.0, {0.7–0.8})	91.9	(0.0, {1.8–2.0})
RAOA	100	({0.5–1.0}, 0.9), ({0.6–1.0}, 1.0)	100	({0.7–1.0}, 0.9)
$d = 8$	96.7	(0.0, 0.8)	86.7	(0.0, {0.7–0.9}), (0.0, {1.1–1.5})
RAHC	100	({0.4–0.5}, 0.8), ({0.6–1.0}, 0.6), (0.9, 0.7)	100	(0.9, 1.4)
$d = 36$	98.0	(0.0, 0.5)	98.0	(0.0, 0.9)

Table 5
Comparative Performance Analysis of Crisp and Fuzzy Equivalence Classes Using LOOCV.

Microarray	Quantitative	Crisp classes: D1		Crisp classes: D2		Fuzzy classes	
data set	measures	Accuracy	Genes	Accuracy	Genes	Accuracy	Genes
Breast	SVM	100	18	100	8	100	6
	K-NN	100	45	98.0	27	100	6
Leukemia	SVM	98.6	36	100	14	100	4
	K-NN	100	50	100	37	100	3
Colon	SVM	87.1	5	85.5	33	90.3	35
	K-NN	85.5	9	77.4	23	90.3	35
Lung	SVM	100	34	100	8	100	9
	K-NN	100	38	99.5	9	100	4
Prostate	SVM	89.7	44	94.9	48	96.3	43
	K-NN	91.2	5	94.9	23	95.6	6
RAOA	SVM	100	3	100	4	100	4
	K-NN	100	3	100	12	100	8
RAHC	SVM	94.0	36	100	12	100	18
	K-NN	100	11	100	12	100	5

Table 6
Classification accuracy of SVM and K-NN rule on full gene set.

Experimental	Methods/	Statistical	Microarray gene expression data sets						
setup	measures	values	Breast	Leukemia	Colon	Lung	Prostate	RAOA	RAHC
LOOCV	SVM	Accuracy	91.8	98.6	82.3	98.9	91.9	70.0	96.0
	K-NN	Accuracy	73.5	76.4	74.2	87.9	74.2	76.7	74.0
10-fold CV	SVM	Mean	89.8	98.8	85.5	98.9	92.7	78.3	94.2
		Std.Dev.	10.3	3.8	13.3	2.2	7.3	18.3	9.2
	K-NN	Mean	76.3	75.0	72.6	89.4	72.1	75.0	71.7
		Std.Dev.	10.4	7.8	14.3	6.3	11.9	22.7	17.6

accuracy. That is, the genes for which the samples from different classes are not well separated will not be selected in the reduced set. On the other hand, the presence of irrelevant, redundant, and insignificant genes in the reduced gene set may degrade the quality of the solution.

From the results reported in Tables 6–9, it is seen that the classification accuracy of the K-NN rule and SVM obtained using the mRMR method and proposed algorithm is always higher than that achieved by the whole gene set for all microarray data sets. On the other hand, out of 28 cases, the Quick Reduct algorithm [35], Roughfication [21], and Discernibility Matrix based approach [37] perform better than the whole gene set in 20, 14, and 15 cases, respectively, while the methods proposed by Valdes and Barton [30] and Fang and Busse [29] achieve in 18 and 12 cases, respectively.

Table 7
Comparative performance analysis of mRMR and MRMS algorithms.

Experimental setup		LOOCV				10-fold CV				
Microarray	Methods/	mRMR		MRMS		mRMR		MRMS		Computed
data set	measures	Accuracy	Genes	Accuracy	Genes	Mean	Std.Dev.	Mean	Std.Dev.	value
Breast	SVM	100	6	100	6	100	0.0	100	0.0	–
$m = 7129$	K-NN	100	4	100	6	100	0.0	100	0.0	–
Leukemia	SVM	100	32	100	4	98.8	3.8	100	0.0	1.00
$m = 7070$	K-NN	98.6	18	100	3	98.8	3.8	100	0.0	1.00
Colon	SVM	88.7	10	90.3	35	87.1	11.8	90.7	9.9	0.74
$m = 2000$	K-NN	90.3	11	90.3	35	90.5	14.6	92.1	10.1	0.29
Lung	SVM	99.5	4	100	9	100	0.0	100	0.0	–
$m = 12{,}533$	K-NN	98.3	6	100	4	98.3	3.6	100	0.0	1.49
Prostate	SVM	94.1	20	96.3	43	93.5	5.9	96.3	6.7	1.01
$m = 12{,}600$	K-NN	93.4	31	95.6	6	92.8	5.6	95.7	4.8	1.26
RAOA	SVM	100	4	100	4	100	0.0	100	0.0	–
$m = 18{,}432$	K-NN	100	3	100	8	100	0.0	100	0.0	–
RAHC	SVM	100	29	100	18	100	0.0	100	0.0	–
$m = 41{,}056$	K-NN	100	11	100	5	100	0.0	100	0.0	–

Table 8
Comparative performance analysis of different rough set based algorithms using LOOCV.

Microarray	Methods/	Fang and Busse		Roughfication		Valdes–Barton		Quick reduct		Discern. matrix		MRMS	
data set	measures	Accuracy	Genes	Accuracy	Genes	Accuracy	Genes	Accuracy	Genes	Accuracy	Genes	Accuracy	Genes
Breast	SVM	73.5	7	77.6	7	81.7	1	85.7	3	71.4	5	100	6
$m = 7129$	K-NN	71.4	6	79.6	49	89.8	1	83.7	2	73.5	3	100	6
Leukemia	SVM	86.1	6	84.7	16	93.1	1	100	3	95.8	4	100	4
$m = 7070$	K-NN	79.2	6	80.6	7	93.1	1	98.6	2	91.7	1	100	3
Colon	SVM	64.5	1	85.5	241	85.5	1	80.7	2	83.9	4	90.3	35
$m = 2000$	K-NN	61.3	2	80.7	6	85.5	1	80.7	3	82.3	5	90.3	35
Lung	SVM	99.5	4	*	*	97.3	1	99.5	3	99.5	5	100	9
$m = 12{,}533$	K-NN	98.9	3	*	*	97.2	1	99.5	3	93.9	5	100	4
Prostate	SVM	56.6	3	*	*	74.3	7	84.6	4	75.0	10	96.3	43
$m = 12{,}600$	K-NN	78.7	4	*	*	84.6	1	88.2	4	75.0	10	95.6	6
RAOA	SVM	70.0	1	86.7	1	83.3	1	86.7	1	76.7	4	100	4
$m = 18{,}432$	K-NN	73.3	1	93.3	3	90.0	1	90.0	2	86.7	3	100	8
RAHC	SVM	70.0	1	82.0	6	86	1	70.0	1	*	*	100	18
$m = 41{,}056$	K-NN	80.0	1	84.0	8	84.0	1	84.0	3	*	*	100	5

6.6.2. Comparative performance of mRMR and MRMS

To compare the performance of the proposed MRMS method with that of the mRMR method [8], extensive experimentation is carried out on seven microarray data sets. Both LOOCV and 10-fold CV are performed on each gene expression data sets.

Table 7 presents the classification accuracy of both SVM and K-NN rule for the MRMS and mRMR methods, along with the computed test statistic values for 10-fold CV. From the results reported in Table 7, it is seen that the proposed MRMS algorithm selects a set of relevant and significant genes from the whole gene set having highest classification accuracy of both SVM and K-NN rule in all the cases. Out of total 28 cases, the proposed method achieves 100% classification accuracy in 20 cases, while the mRMR method attains this accuracy in 14 cases. However, the mRMR method attains same K-NN accuracy for breast cancer, colon cancer, and RAOA data set as that of the proposed MRMS method with lesser number of genes. Also, the computed test statistic values indicate that although the MRMS method performs better than the mRMR method, the results are not significantly better as all the computed values are less than 1.81, which is the tabled value at an error probability level of 0.05.

6.6.3. Performance of different rough set based algorithms

Finally, Tables 8 and 9 compare the best performance of different existing rough set based feature selection algorithms with that of the proposed MRMS algorithm. While Table 8 presents the classification accuracy of both SVM and K-NN rule using the LOOCV, Table 9 depicts that using 10-fold CV.

From the results reported in Table 8, it is seen that the proposed MRMS algorithm achieves highest classification accuracy of both SVM and K-NN rule in all the cases. Out of total 14 cases, the proposed method achieves 100% classification accuracy in 10 cases, while the Quick Reduct algorithm [35] attains this accuracy in only one case. However, the Quick Reduct algorithm attains same SVM accuracy for leukemia data set as that of the proposed MRMS method with lesser number of genes.

Table 9
Comparative test statistic analysis of different rough set based algorithms using 10-fold CV.

Data	Statistical	MRMS		Fang and Busse		Roughfication		Valdes–Barton		Quick Reduct		Discern. matrix	
sets	values	SVM	K-NN	SVM	K-NN	SVM	K-NN	SVM	K-NN	SVM	K-NN	SVM	K-NN
Breast	Mean	100.0	100.0	77.7	73.0	75.8	79.2	85.3	89.8	85.8	84.2	73.0	74.2
	Std.Dev.	0.0	0.0	14.2	16.3	20.3	17.2	18.4	13.7	13.0	14.2	18.8	21.1
	Computed			4.97	5.24	3.77	3.84	2.53	2.35	3.44	3.52	4.53	3.87
Leukemia	Mean	100.0	100.0	89.1	80.2	85.5	77.3	93.4	93.4	100.0	97.3	96.3	91.7
	Std.Dev.	0.0	0.0	11.9	10.0	12.2	10.0	6.6	6.6	0.0	5.4	8.0	8.9
	Computed			2.88	6.26	3.76	7.13	3.15	3.15	–	1.58	1.48	2.96
Colon	Mean	90.7	92.1	64.8	63.1	85.9	81.2	87.4	85.7	82.6	80.9	84.3	82.6
	Std.Dev.	9.9	10.2	3.8	9.4	12.1	12.7	13.9	13.3	14.4	13.8	13.9	12.9
	Computed			7.70	6.63	0.96	2.13	0.62	1.22	1.46	2.07	1.19	1.83
Lung	Mean	100.0	100.0	100.0	98.9	*	*	97.2	97.2	99.4	99.4	97.2	94.5
	Std.Dev.	0.0	0.0	0.0	2.2	*	*	3.7	2.8	1.7	1.7	4.5	6.1
	Computed			–	1.58	*	*	2.36	3.16	1.06	1.06	1.96	2.87
Prostate	Mean	96.3	95.7	55.9	80.2	*	*	73.6	84.5	83.7	86.1	75.1	74.1
	Std.Dev.	6.7	4.8	2.4	8.1	*	*	13.3	5.9	10.3	8.2	10.1	6.3
	Computed			18.07	5.21	*	*	4.84	4.63	3.26	3.18	5.57	8.65
RAOA	Mean	100.0	100.0	70.8	75.0	88.3	93.3	86.7	90.0	90.8	93.3	80.0	86.7
	Std.Dev.	0.0	0.0	10.0	17.1	18.3	24.1	16.3	15.3	14.2	13.3	16.3	16.3
	Computed			9.20	4.63	2.01	0.88	2.58	2.07	2.05	1.58	3.87	2.58
RAHC	Mean	100.0	100.0	70.8	77.5	83.3	84.2	85.0	84.2	70.8	87.5	*	*
	Std.Dev.	0.0	0.0	4.2	17.9	11.8	15.6	12.8	11.5	4.2	13.0	*	*
	Computed			22.12	3.98	4.47	3.22	3.71	4.37	22.12	3.03	*	*

Similarly, the results using 10-fold CV reported in Table 9 show that the proposed MRMS methods attains 100% classification accuracy in 10 cases, while both Quick Reduct algorithm [35] and the method proposed by Fang and Busse [29] attains this accuracy in only one case. Also, the performance of the MRMS method is always better than that of any existing rough set based algorithms. Out of 70 comparisons, the proposed method is found to provide significantly better results in 56 comparisons. Other 14 cases, the performance of the MRMS method is found to be better, but not significantly. The better performance of the proposed gene selection algorithm is achieved due to the fact that it can identify relevant and significant genes from microarray data sets more accurately than the existing rough set based algorithms.

7. Conclusion and future works

The main contribution of this paper is threefold, namely,

(1) development of a new feature selection method based on the rough set theory;
(2) application of the proposed method in identifying discriminative and significant genes from high-dimensional microarray gene expression data sets; and
(3) compare the performance of the proposed method and some existing methods using the predictive accuracy of K-nearest neighbor rule and support vector machine.

For five cancer and two arthritis microarray data sets, significantly better results are found for the proposed method compared to existing rough set based methods. All the results reported in this paper demonstrate the feasibility and effectiveness of the proposed feature selection method. It is capable of identifying discriminative and significant genes that may contribute to revealing underlying class structures, providing a useful tool for the exploratory analysis of biological data.

The results obtained on different microarray data sets demonstrate that the proposed method can bring a remarkable improvement on gene selection problem. The proposed method is only used for selection of genes from high dimensional microarray data sets. In future, this method will be extended to other feature selection tasks and further its merits and limitations will be evaluated. It will also be combined with fuzzy-rough sets [26,32,54–57] and neighborhood rough sets [58] in near future to deal with numerical features directly without discretization. A method will be developed based on some quantitative measures to find out the optimum values of different parameters. In order to address the problem of multiplicity of marker genes, a detailed analysis of the biological relevance of the selected genes will be conducted in future. The gene interactions will be studied in detail to see whether incorporation of gene interaction information can improve the diagnostic test.

References

[1] 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, E.S. Lander, Molecular classification of cancer: class discovery and class prediction by gene expression monitoring, Science 286 (5439) (1999) 531–537.
[2] E. Domany, Cluster analysis of gene expression data, Journal of Statistical Physics 110 (3–6) (2003) 1117–1139.

[3] J.G. Liao, K.-V. Chin, Logistic regression for disease classification using microarray data: model selection in a large p and small n case, Bioinformatics 23 (15) (2007) 1945–1951.
[4] F. Napolitano, G. Raiconi, R. Tagliaferri, A. Ciaramella, A. Staiano, G. Miele, Clustering and visualization approaches for human cell cycle gene expression data analysis, International Journal of Approximate Reasoning 47 (2008) 70–84.
[5] P.A. Devijver, J. Kittler, Pattern Recognition: A Statistical Approach, Prentice-Hall, Englewood Cliffs, 1982.
[6] D. Koller, M. Sahami, Toward optimal feature selection, in: Proceedings of the International Conference on Machine Learning, 1996, pp. 284–292.
[7] R. Kohavi, G.H. John, Wrappers for feature subset selection, Artificial Intelligence 97 (1–2) (1997) 273–324.
[8] C. Ding, H. Peng, Minimum redundancy feature selection from microarray gene expression data, in: Proceedings of the Computational Systems Bioinformatics, 2003, pp. 523–528.
[9] J. Li, H. Su, H. Chen, B.W. Futscher, Optimal search-based gene subset selection for gene array cancer classification, IEEE Transactions on Information Technology in Biomedicine 11 (4) (2007) 398–405.
[10] H. Peng, F. Long, C. Ding, Feature selection based on mutual information: criteria of Max-Dependency, Max-Relevance, and Min-Redundancy, IEEE Transactions on Pattern Analysis and Machine Intelligence 27 (8) (2005) 1226–1238.
[11] X. Liu, A. Krishnan, A. Mondry, An entropy based gene selection method for cancer classification using microarray data, BMC Bioinformatics 6 (76) (2005) 1–14.
[12] I. Inza, B. Sierra, P. Larranaga, R. Blanco, Gene selection by sequential wrapper approaches in microarray cancer class prediction, Journal of Intelligent and Fuzzy Systems 12 (2002) 25–33.
[13] R. Blanco, P. Larranaga, I. Inza, B. Sierra, Gene selection for cancer classification using wrapper approaches, International Journal of Pattern Recognition and Artificial Intelligence 18 (8) (2004) 1373–1390.
[14] L. Wang, J. Zhu, H. Zou, Hybrid huberized support vector machines for microarray classification and gene selection, Bioinformatics 24 (3) (2008) 412–419.
[15] J.-H. Hong, S.-B. Cho, Gene boosting for cancer classification based on gene expression profiles, Pattern Recognition 42 (9) (2009) 1761–1767.
[16] P. Maji, f-Information measures for efficient selection of discriminative genes from microarray data, IEEE Transactions on Biomedical Engineering 56 (4) (2009) 1063–1069.
[17] D. Jiang, C. Tang, A. Zhang, Cluster analysis for gene expression data: a survey, IEEE Transactions on Knowledge and Data Engineering 16 (11) (2004) 1370–1386.
[18] L.J. Heyer, S. Kruglyak, S. Yooseph, Exploring expression data: identification and analysis of coexpressed genes, Genome Research 9 (11) (1999) 1106–1115.
[19] A. Gruzdz, A. Ihnatowicz, D. Slezak, Interactive gene clustering – a case study of breast cancer microarray data, Information Systems Frontiers 8 (2006) 21–27.
[20] D. Slezak, Rough sets and few-objects-many-attributes problem: the case study of analysis of gene expression data sets, in: Proceedings of the Frontiers in the Convergence of Bioscience and Information Technologies 2007, pp. 233–240.
[21] D. Slezak, J. Wroblewski, Roughfication of numeric decision tables: the case study of gene expression data, Proceedings of the Second International Conference on Rough Sets and Knowledge Technology, Springer, Berlin/Heidelberg, 2007, pp. 316–323.
[22] Z. Pawlak, Rough Sets, Theoretical Aspects of Resoning About Data, Kluwer, Dordrecht, The Netherlands, 1991.
[23] A. Skowron, R.W. Swiniarski, P. Synak, Approximation spaces and information granulation, Transactions on Rough Sets 3 (2005) 175–189.
[24] C. Cornelis, R. Jensen, G.H. Martin, D. Slezak, Attribute selection with fuzzy decision reducts, Information Sciences 180 (2010) 209–224.
[25] R. Jensen, Q. Shen, Fuzzy-rough attribute reduction with application to web categorization, Fuzzy Sets and Systems 141 (2004) 469–485.
[26] W. Wu, W. Zhang, Constructive and axiomatic approaches of fuzzy approximation operators, Information Sciences 159 (2004) 233–254.
[27] Y. Kudo, T. Murai, S. Akama, A granularity-based framework of deduction, induction, and abduction, International Journal of Approximate Reasoning 50 (8) (2009) 1215–1226.
[28] Y. Qian, J. Liang, C. Dang, Knowledge structure, knowledge granulation and knowledge distance in a knowledge base, International Journal of Approximate Reasoning 50 (1) (2009) 174–188.
[29] J. Fang, J.W.G. Busse, Mining of MicroRNA expression data – a rough set approach, in: Proceedings of the First International Conference on Rough Sets and Knowledge Technology, Springer, Berlin/Heidelberg, 2006, pp. 758–765.
[30] J.J. Valdes, A.J. Barton, Relevant attribute discovery in high dimensional data: application to breast cancer gene expressions, in: Proceedings of the First International Conference on Rough Sets and Knowledge Technology, Springer, Berlin/Heidelberg, 2006, pp. 482–489.
[31] Q. Shen, A. Chouchoulas, Combining rough sets and data-driven fuzzy learning for generation of classification rules, Pattern Recognition 32 (12) (1999) 2073–2076.
[32] R. Jensen, Q. Shen, Semantics-preserving dimensionality reduction: rough and fuzzy-rough-based approach, IEEE Transactions on Knowledge and Data Engineering 16 (12) (2004) 1457–1471.
[33] K.H. Chen, Z.W. Ras, A. Skowron, Attributes and rough properties in information systems, International Journal of Approximate Reasoning 2 (1988) 365–376.
[34] D. Yamaguchi, Attribute dependency functions considering data efficiency, International Journal of Approximate Reasoning 51 (2009) 89–98.
[35] A. Chouchoulas, Q. Shen, Rough set-aided keyword reduction for text categorisation, Applied Artificial Intelligence 15 (2001) 843–873.
[36] J. Komorowski, Z. Pawlak, L. Polkowski, A. Skowron, Rough sets: a tutorial, in: S. Pal, A. Skowron (Eds.), Rough-Fuzzy Hybridization: A New Trend in Decision Making, Springer Verlag, Singapore, 1999, pp. 3–98.
[37] A. Skowron, C. Rauszer, The discernibility matrices and functions in information systems, in: R. Slowinski (Ed.), Intelligent Decision Support, Kluwer Academic Publishers, Dordrecht, 1992, pp. 331–362.
[38] J. Bazan, A. Skowron, P. Synak, Dynamic reducts as a tool for extracting laws from decision tables, in: Z.W. Ras, M. Zemankova (Eds.), Proceedings of the 8th Symposium on Methodologies for Intelligent Systems, Lecture Notes in Artificial Intelligence, vol. 869, Springer-Verlag, 1994, pp. 346–355.
[39] W. Ziarko, Variable precision rough set model, Journal of Computer and System Sciences 46 (1993) 39–59.
[40] G. Xie, J. Zhang, K. Lai, L. Yu, Variable precision rough set for group decision-making: an application, International Journal of Approximate Reasoning 49 (2008) 331–343.
[41] M. Inuiguchi, Y. Yoshioka, Y. Kusunoki, Variable-precision dominance-based rough set approach and attribute reduction, International Journal of Approximate Reasoning 50 (2009) 1199–1214.
[42] D. Kim, Data classification based on tolerant rough set, Pattern Recognition 34 (8) (2001) 1613–1624.
[43] N.M. Parthalain, Q. Shen, Exploring the boundary region of tolerance rough sets for feature selection, Pattern Recognition 42 (5) (2009) 655–667.
[44] J. Yao, Y. Yao, W. Ziarko, Probabilistic rough sets: approximations, decision-makings, and applications, International Journal of Approximate Reasoning 49 (2) (2008) 253–254.
[45] Y. Yao, Probabilistic rough set approximations, International Journal of Approximate Reasoning 49 (2) (2008) 255–271.
[46] W. Ziarko, Probabilistic approach to rough sets, International Journal of Approximate Reasoning 49 (2) (2008) 272–284.
[47] M. Modrzejewski, Feature selection using rough sets theory, in: Proceedings of the 11th International Conference on Machine Learning, 1993, pp. 213–226.
[48] N. Zhong, J. Dong, S. Ohsuga, Using rough sets with heuristics for feature selection, Journal of Intelligent Information Systems 16 (2001) 199–214.
[49] A. Bjorvand, J. Komorowski, Practical applications of genetic algorithms for efficient reduct computation, in: Proceedings of the 15th IMACS World Congress on Scientific Computation, Modeling and Applied Mathematics, vol. 4, 1997, pp. 601–606.
[50] D. Ślęzak, Approximate reducts in decision tables, in: Proceedings of the 6th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU -96), 1996, pp. 1159–1164.
[51] J. Wróblewski, Finding minimal reducts using genetic algorithms, in: Proceedings of the 2nd Annual Joint Conference on Information Sciences, 1995, pp. 186–189.
[52] J. Han, M. Kamber, Data Mining, Concepts and Techniques, Morgan Kaufman Publishers, 2001.
[53] M.J. Beynon, Stability of continuous value discretisation: an application within rough set theory, International Journal of Approximate Reasoning 35 (2004) 29–53.

[54] Q. Hu, Z. Xie, D. Yu, Hybrid attribute reduction based on a novel fuzzy-rough model and information granulation, Pattern Recognition 40 (2007) 3577–3594.
[55] S. Zhao, E.C.C. Tsang, D. Chen, The model of fuzzy variable precision rough sets, IEEE Transactions on Fuzzy Systems 17 (2009) 451–467.
[56] R. Jensen, Q. Shen, Fuzzy-rough sets assisted attribute selection, IEEE Transactions on Fuzzy Systems 15 (2007) 73–89.
[57] Q. Hu, L. Zhang, D. Chen, W. Pedrycz, D. Yu, Gaussian kernel based fuzzy rough sets: model, uncertainty measures and applications, International Journal of Approximate Reasoning 51 (2010) 453–471.
[58] Q. Hu, D. Yu, J. Liu, C. Wu, Neighborhood rough set based heterogeneous feature subset selection, Information Sciences 178 (2008) 3577–3594.
[59] P. Maji, S.K. Pal, Fuzzy-rough sets for information measures and selection of relevant genes from microarray data, IEEE Transactions on System, Man and Cybernetics, Part B, Cybernetics 40 (3) (2010) 741–752.
[60] P. Maji, S.K. Pal, Feature selection using f-information measures in fuzzy approximation spaces, IEEE Transactions on Knowledge and Data Engineering 22 (6) (2010) 854–867.
[61] S.K. Pal, S. Mitra, Neuro-Fuzzy Pattern Recognition: Methods in Soft Computing, Wiley, New York, 1999.
[62] M. West, C. Blanchette, H. Dressman, E. Huang, S. Ishida, R. Spang, H. Zuzan, J.A. Olson, J.R. Marks, J.R. Nevins, Predicting the clinical status of human breast cancer by using gene expression profiles, Proceedings of the National Academy of Science, USA 98 (20) (2001) 11462–11467.
[63] U. Alon, N. Barkai, D.A. Notterman, K. Gish, S. Ybarra, D. Mack, A.J. Levine, Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays, Proceedings of the National Academy of Science, USA 96 (12) (1999) 6745–6750.
[64] G.J. Gordon, R.V. Jensen, L.-L. Hsiao, S.R. Gullans, J.E. Blumenstock, S. Ramaswamy, W.G. Richards, D.J. Sugarbaker, R. Bueno, Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma, Cancer Research 62 (2002) 4963–4967.
[65] D. Singh, P.G. Febbo, K. Ross, D.G. Jackson, J. Manola, C. Ladd, P. Tamayo, A.A. Renshaw, A.V. D'Amico, J.P. Richie, E.S. Lander, M. Loda, P.W. Kantoff, T.R. Golub, W.R. Sellers, Gene expression correlates of clinical prostate cancer behavior, Cancer Research 1 (2002) 203–209.
[66] T.C.T.M. van der Pouw Kraan, F.A. van Gaalen, P.V. Kasperkovitz, N.L. Verbeet, T.J.M. Smeets, M.C. Kraan, M. Fero, P.-P. Tak, T.W.J. Huizinga, E. Pieterman, F.C. Breedveld, A.A. Alizadeh, C.L. Verweij, Rheumatoid arthritis is a heterogeneous disease: evidence for differences in the activation of the STAT-1 pathway between rheumatoid tissues, Arthritis and Rheumatism 48 (8) (2003) 2132–2145.
[67] T.C.T.M. van der Pouw Kraan, C.A. Wijbrandts, L.G.M. van Baarsen, A.E. Voskuyl, F. Rustenburg, J.M. Baggen, S.M. Ibrahim, M. Fero, B.A.C. Dijkmans, P.P. Tak, C.L. Verweij, Rheumatoid arthritis subtypes identified by genomic profiling of peripheral blood cells: assignment of a Type I interferon signature in a subpopulation of pateints, Annals of the Rheumatic Diseases 66 (2007) 1008–1014.
[68] V. Vapnik, The Nature of Statistical Learning Theory, Springer-Verlag, New York, 1995.
[69] R.O. Duda, P.E. Hart, D.G. Stork, Pattern Classification and Scene Analysis, John Wiley and Sons, New York, 1999.
[70] A. Aspin, Tables for use in comparisons whose accuracy involves two variances, Biometrika 36 (1949) 245–271.