Scientific Research and Essays Vol. 7(14), pp. 1512-1520, 16 April, 2012
Available online at http://www.academicjournals.org/SRE
DOI: 10.5897/SRE11.2171
ISSN 1992-2248 ©2012 Academic Journals

Full Length Research Paper

Fuzzy-rough feature selection and a fuzzy 2-level complementary approach for classification of gene expression data

Zahra Shaeiri[1], Reza Ghaderi[1]* and Ali Hojjatoleslami[2]

[1]Department of Electrical and Computer Engineering, Babol University of technology, Babol, Iran.
[2]School of Bioscience, University of Kent, Canterbury, Kent, Brussels, Belgium.

Accepted 30 March, 2012

Classification of gene expression data is an important issue in medical diagnosis of disease such as cancer. In this paper first Fuzzy-Rough Set theory is established to select relevant features for classification. This will be followed by proposing a new fuzzy 2-level complementary learning method. The Fuzzy-Rough Set is a mathematical tool which encapsulates the relevant but distinct concepts of fuzziness and indiscernibility. These are caused due to uncertainties in knowledge or datasets. Thus a feature selection using this tool is designed to handle two complementary kinds of uncertainties and to increase the accuracy of the outcome. Complementary learning mechanism, on the other hand, has significant performance because it is responsible for human pattern recognition whose is effective in the learning stage and the problem solving. The proposed classification system works in two levels of different accuracies. If the first level fails to process the sample, the second level would handle. A simulation is carried out using some published datasets. The performance of the proposed classification method by means of achieving an excellent accuracy rate of the classification will be shown significantly with respect to some recently proposed methods.

Key words: Gene expression data analysis, fuzzy-rough set feature selection, complementary learning method, hierarchical fuzzy system.

INTRODUCTION

The cDNA microarray technology has caused a major change in the field of life science. It introduces a new experimental technique in which it is possible to monitor the expression levels of all known genes in a specific organism (Kohane, 2003; Xing, 2006). Nowadays some datasets which contain the gene expression profile data are available [gene expression data, 2012]. The analysis of these datasets is important to diagnose disease and therefore it requires advanced data analysis methods.

There are some conventional methods such as linear discriminant analysis (Xiong, 2001), nearest neighbor (Dudoit, 2000), and advanced methods such as fuzzy logic (Ohno-Machado, 2002), neural networks (han, 2001), decision trees (Cai, 2000) and support vector machines (Duan, 2005; Guyon, 2002) Although, many classification methods are available, in the gene expression data analysis conventional approaches encounter the problem of curse of dimensionality. Typically, the gene expression datasets contain a huge number of genes (tens of thousands), and a small number of tuples (tens or few hundred). Only a fraction of these huge amounts of genes are necessary from the

*Corresponding author. E-mail: r_ghaderi@nit.ac.ir.

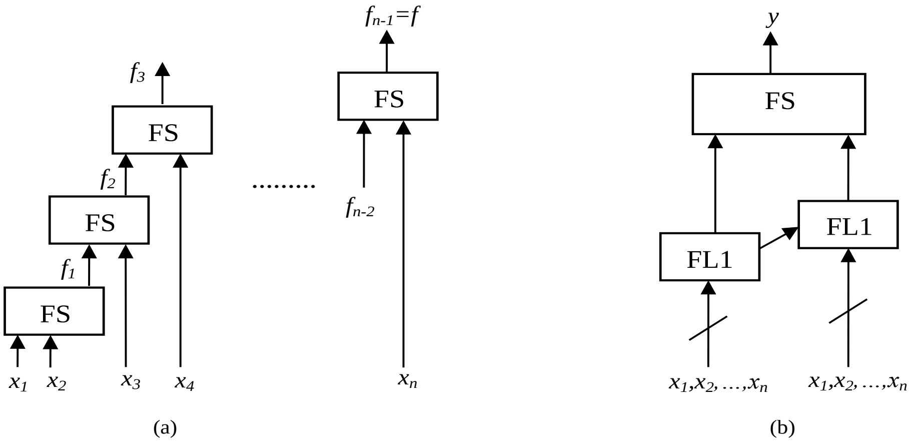

Figure 1(a). A conventional hierarchical fuzzy system (b). The Proposed hierarchical fuzzy system.

clinical or biological view point. Therefore, selecting the relevant genes together with less losing the information is crucial. There are some feature selection methods which fall into two categories (Kohavi, 1997; Langley, 1994): filtering methods and wrapper techniques. In the filtering methods, genes are selected based on their dependency to certain classes such as statistical test scores (t-test) (Ding, 2002), Pearson correlation coefficient (PCC) and SNR-evolving classification function (ECF) (Guh, 2004). Unlike the filtering procedure, wrapper methods apply to the estimated accuracy of a specific classifier to evaluate candidate subsets of features. In (Guyon, 2002) they proposed a support vector machine recursive feature elimination (SVM-RFE) approach to select genes.

Rough Set theory was proposed by pawlak in 1982 (Pawlak, 1982). Thereafter, it has become a topic with great interests and has been applied to many domains of feature selection and classification task. In contrast to other approaches in the feature selection context (Dash, 1997; Devijver, 1982; Kira, 1992), Rough Set theory needs no further information (the thresholds or the expert knowledge) except the dataset itself. In the Rough Set theory, the hidden facts in the dataset are analyzed and accordingly a minor knowledge representation will be at last achieved (Jensen, 2007). Although the Rough Set theory has led to good results only for discrete datasets, it shows no satisfactory results facing with real valued datasets. This was a significant shortcoming, while many real world problems are real valued ones. One of the best extensions of the Rough Set theory is the Fuzzy-Rough Set theory which was primarily proposed by Shen and Jensen (Jensen, 2001), (Jensen, 2004). Their work performed well on many practical applications.

In this paper we primarily exploit the feature selection capability of the Fuzzy-Rough Set theory. The work will be followed by removing some redundant genes. Accordingly a 2-level complementary fuzzy classifier for classification of cancer-type samples and non-cancer ones will be proposed. This classifier uses both the learning and the complementary learning mechanisms. The first one is a plausible reasoning whilst the latter underlies on human pattern recognition by means of assessing positive and negative samples apart from each other in the brain areas. Visual processing, from the retina through the inferotemporal and parietal cortices, provides excellent examples of complementary processing systems in the brain. Each brain area is registered by some knowledge from certain object (Tan, 2008). When any object is presented, the associated knowledge with the activated object will be fired. The other part of the knowledge is inhibited. This is the basis of the complementary mechanism of human brain which is believed to be capable of enhancing the performance of the pattern recognition systems.

The proposed classifier consists of a two hierarchy levels of accuracy. The first Level has less accuracy than the 2nd one. When more classification resolution and accuracy are needed the 2nd level will be activated to decide about the output of the classification task. As shown in Figure 1(b), both of two levels are fed by all attributes. The main difference between two levels is the required accuracy for the class discovery purpose. In Figure 1(a), a common Hierarchical Fuzzy System (HFS) is illustrated. A common hierarchical fuzzy system is composed of some low dimensional fuzzy systems in a hierarchical order. It means that the output of one layer is an input to the next layer. This kind of structure decreases the number of fuzzy rules in price of

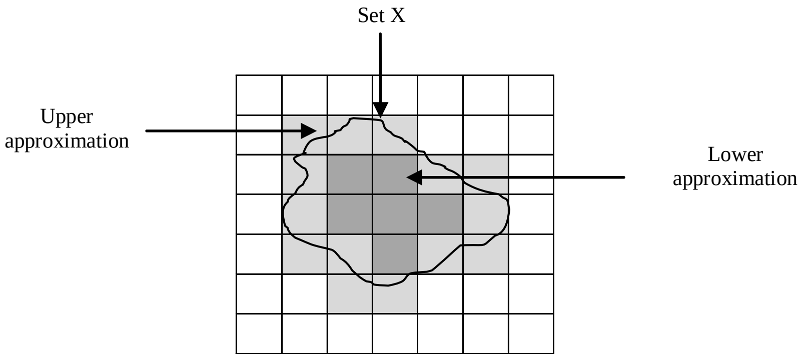

Figure 2. S-Lower and S-Upper approximations of set *X*.

increasing the computational efficiency. However an important drawback of such structures is that the retrieval of a physical meaning from the fuzzy rules in the middle of the hierarchy is not an easy task. As shown in Figure 1(b), since the intermediate rules are vanished, the proposed hierarchical fuzzy system is more comprehensible than the conventional approaches like in (Joo, 2005). Functionally modeling of the human psychological abilities such as hierarchical and complementary methods can reduce the complexity of the system and enhance the performances of the classifier at the same time.

Attribute selection

The Fuzzy-Rough set

Several applications process data in the form of real-valued vectors. The gene expression data analysis, text classification, and bookmark categorization are such applications. If these vectors are of the high dimensionality, the processing becomes infeasible. Therefore a technique must be used to discover the data dependencies and reduce the dimensionality. Beside, after dimensionality reduction, meanings of features must be preserved for the next stages. Semantic-destroying dimensionality reduction (feature extraction) irreversibly transforms data, whereas semantic-preserving ones (feature selection) attempt to retain the meaning of the original feature set. Fuzzy-Rough Set theory is a semantic-preserving technique which provides the means of data reduction for crisp and real-valued datasets. It utilizes the extent to which values are similar. Fuzzy-Rough Set feature selection method is based on the concept of indiscernibility relation which partitions the domain. Given a rough concept (set X in Figure 2); the purpose is to approximate it by constructing two exact concepts. These two concepts are the lower and upper approximations, which are determined by the indiscernibility relation. Each square in Figure 2 represents an equivalence class, generated by indiscernibility between object values. From these equivalence classes lower and upper approximations of set X can be constructed. Equivalence classes lying within X belong to the lower approximation which is a set of objects definitely belonging to the rough concept X. equivalence classes within X and along its border form the upper approximation. Objects in this region can only be said to possibly belong to the concept X (Jensen, 2001). In the following some basic definitions of Fuzzy-Rough Sets and the reduction method are summarized.

Definition 1: Consider an information system as $I=(U,A)$ in which *U* is a universe and *A* is a set of features, *a*. Both *U* and *A* are non-empty and finite sets. $a:U\rightarrow V_a$, where V_a is a value that feature *a* may take. An information system can be viewed as a decision table$\left(A=a_C\cup a_D\right)$. a_C is a conditional feature and a_D is a decision feature. a_C and a_D are crisp in Rough Set and fuzzy in Fuzzy-Rough Set theory.

Definition 2: In the Rough Set theory for any subset of conditional features $S\subseteq A$ an equivalence relation or *S*-indiscernibility relation *IND(S)* is defined as follows:

$$IND(S)=\left\{(x,y)\in U^2 \mid \forall a\in S, a(x)=a(y)\right\} \qquad (1)$$

IND(S) partitions the universe *U* into some crisp equivalence classes (*U/S*). $[x_i]_S$ is the set of universe including x_i that are indiscernible in *S*:

$$[x_i]_S=\left\{x_j\in U:(x_i,x_j)\in IND(S)\right\}, \qquad x_i\in U \qquad (2)$$

By the extension of the crisp equivalence classes a fuzzy equivalence relation *S* can be made on the universe, which determines the extent to which two elements are similar in *S*. Fuzzy equivalence class $[x]_S$ for objects close to *x* can be defined as:

$$\mu_{[x]_S}(y) = \mu_S(x,y) \tag{3}$$

Definition 3: Let $X \subseteq U$. Crisp *S*-lower and *S*-upper approximations of *X*, which are the least and the greatest definable set contained in *X*, can be approximated as follows (Figure 2):

$$\begin{aligned} \underline{S}X &= \{x \in U : [x]_S \subseteq X\} \\ \bar{S}X &= \{x \in U : [x]_S \bigcap X \neq \phi\} \end{aligned} \tag{4}$$

Let *S* and *O* be two equivalence relations over *U*. Region of objects that can be classified to classes of *U/O* using only information from feature set *S* are so called positive region, $POS_S(O)$. Region of objects where can be possibly classified in this way is called boundary region, $BND_S(O)$, and negative region includes objects which cannot be classified into classes of *U/O* using information from feature set *S*. These regions are defined as follows:

$$\begin{aligned} POS_S(O) &= \bigcup_{X \in U/O} \underline{S}X \\ NEG_S(O) &= U - \bigcup_{X \in U/O} \bar{S}X \\ BND_S(O) &= \bigcup_{X \in U/O} \bar{S}X - \bigcup_{X \in U/O} \underline{S}X \end{aligned} \tag{5}$$

The positive region in the crisp Rough Set theory is defined as the union of the lower approximations. In the Rough Set feature selection method positive region is used. If $POS_S(O) = POS_{S\backslash\{b\}}(O)$, then feature *b* is said to be dispensible in *S*, otherwise *b* is indispensible in $S\ (b \in S(\subseteq a_C))$.

Similarly fuzzy *S*-lower $\mu_{\underline{S}X}(x)$ and *S*-upper $\mu_{\bar{S}X}(x)$ approximations of *X* are defined as:

$$\begin{aligned} \mu_{\underline{S}X}(x) &= \sup_{F \in U/S} \min(\mu_F(x), \inf_{y \in U} \max\{1 - \mu_F(y), \mu_X(y)\}) \\ \mu_{\bar{S}X}(x) &= \sup_{F \in U/S} \min(\mu_F(x), \sup_{y \in U} \min\{\mu_F(y), \mu_X(y)\}) \end{aligned} \tag{6}$$

Where *S* is an equivalence class, *X* is the concept which is needed to be approximated, and *F* is a fuzzy equivalence class belonging to *U/S*. The fuzzy positive region can be defined as a membership degree by the extension principal.

$$\mu_{POS_S(Q)}(x) = \sup_{X \in U/Q} \mu_{\underline{P}X}(x) \tag{7}$$

Definition 4: A set of features *O*, is completely depends on another set of features *S*, is denoted by: $S \Rightarrow O$, if all feature values of *O* are uniquely determined by feature values from *S*. If functional dependencies were detected then *O* is completely depends on *S*. In the Rough Set theory the degree of dependency is defined as follows: Given two sets of features $S, O \subseteq A$, *O* depends on *S* in a degree *k*, ($0 \leq k \leq 1$), is denoted by $S \Rightarrow_k O$, if:

$$k = \gamma_S(O) = \frac{|POS_S(O)|}{|U|} \tag{8}$$

$k = 1$ means that *O* completely depends on *S*, $k = 0$ means that *O* does not depend on *S* and $0 < k < 1$ means that *O* partially depends on *S*. using the definition of the fuzzy positive region the fuzzy dependency function is defined as:

$$\gamma_S{}'(Q) = \frac{|\mu_{POS_S(O)}(x)|}{|U|} = \frac{\sum_{x \in U} \mu_{POS_S(O)}(x)}{|U|} \tag{9}$$

Similar to the crisp Rough Sets the dependency of *O* on *S* is a proportion of observations that are discernible out of the whole dataset.

Fuzzy-Rough feature reduction

In feature reduction process, the significance of a feature is evaluated by calculating the change in the dependency function when a feature is removed from the feature set. A feature is more significant if the change in the dependency function is higher. The concept of Reduct, *R* is introduced, which is a minimal subset of initial feature set *C* such that $\gamma_R{}'(Q) = \gamma_C{}'(Q)$. The fuzzy QUICKREDUCT algorithm (Jensen, 2001; Jensen, 2004) is used to generate the minimal Reduct. It begins with an empty set and adds in turn those features which increase the dependency function until it produces its maximum possible value for the dataset.

Hierarchical organization

Many of real-world problems are of the hierarchically

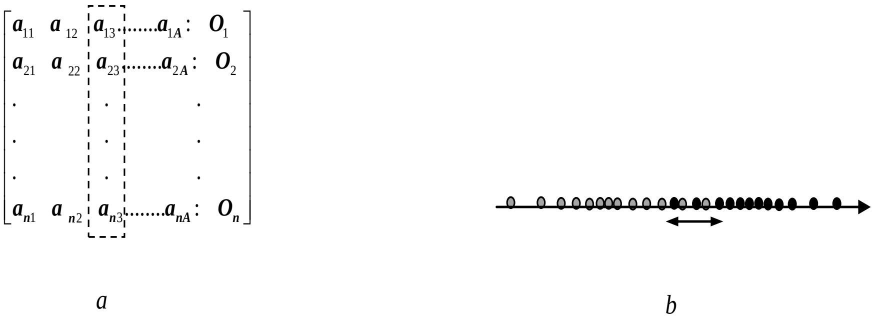

Figure 3(a). A typical gene expression dataset, rows are samples and columns are features or genes (b). Distribution of samples for one gene, gray points are non-cancer samples and black ones are cancer samples.

structured. In fact human being uses a hierarchical representation for organizing his/her knowledge. From behavioral point of view, the hierarchical representation of knowledge in human being allows complex learning problems to be solved by dividing the initial problem in to a set of simpler sub-problems. Thus, facing with a highly organized learning problem, human being concise and acquires the knowledge in a hierarchical structure which itself may contain many levels of hierarchical rules. The rules at higher levels are in relationship with the rules in the lower levels, and all rules cover the whole patterns of the problem.

The higher the hierarchy, more complex patterns are described. Some portions of human brain structures are also hierarchically structured. Cortical system, motivational control center and tactile shape learning are of brain structures in them some regions are hierarchically above some others (Tan, 2008). This hierarchical coding is claimed to reduce the memory usage. In real-world problems, hierarchical architectures arise when we assume that the data are well described by a multi-resolution model - a model in which regions are divided recursively into a number of sub-regions.

Gene expression data classification problem, which here is a two-way cancer non-cancer diagnosis problem, is somehow a multi-resolution problem. Consider matrix shown in Figure 3(a), as a gene expression dataset in which rows are samples and columns are attributes. The last column is the output which is 0 for the cancer type and 1 for a non-cancer type sample. We depicted distribution of samples for one attribute in Figure 3(b) in which gray points are of the cancer types and black points are of the non-cancer. It can be seen that in the marked region, there are some samples belonging to both types. Since samples in this region have degrees of membership in the positive fuzzy sets which is near or maybe equal to degrees of the membership in the negative fuzzy sets, in this region much more accuracy is needed. In this case having hierarchy of levels in order to effectively distinguish between the two classes might be prominent. However, this fact is a motivation to construct a fuzzy classification system with a hierarchical structure for classifying gene expression dataset. Each linguistic term characterizes a concept. In this work there are 2 levels of hierarchy, a specific low level and a general high level, that is $A=\underline{A}\cup\bar{A}$, where $\underline{A}$ is a specific concept and $\bar{A}$ is a general concept. Since each defined linguistic term by its membership function, can be represented as:

$$\mu_A(x_j)=\begin{cases}\mu_{\bar{A}}(x_j)\\ \mu_{\underline{A}}(x_j)\end{cases} \qquad (10)$$

Formally speaking, we can formulate the hierarchical structure of the fuzzy system as follows: The first level is a common fuzzy system, but in the second level features are ranked based on the discrepancy of the positive samples mean value and the negative ones. Thus, for the second level, we define degree of membership of each attribute as follows:

$$\mu_{\bar{A}_i}(x)=\sum_{m=1}^{i}\mu_{\underline{A}_m}(x_j) \qquad (11)$$

In which $\bar{A}_i$ and $\underline{A}_i$ are the high level and low level concept of the fuzzy set A_i respectively, *i* is the number of features and x_j is *j*th sample.

Complementary learning mechanism

Complementary or positive and negative leaning mechanism is a plausible reasoning and learning method that underlies on human pattern recognition. Human registers different brain areas to recognize different objects. For instance, when a car is presented (positive sample), only brain areas registered for the car recognition (positive rules) are activated, whereas brain areas registered for other objects (negative rules) are concurrently inhibited, and vice versa (Gauthier, 2000), (Shaeiri, 2011). A complementary learning paradigm has some unique characteristics compared with traditional concept learning methods such as AQ_{14}, C4.5 and fuzzy clustering relevant feedback (DeJong, 1993). The complementary learning paradigm can be formulized as follows: For each instance, $A=\{x^1,x^2,x^3,...,x^A\}$ which is a set of input attributes to the system, positive and negative weights should be calculated. A fuzzy set A_C representing a particular concept *C*, the elements in the set *A* will have a unit membership if they belong to the concept *C*, as shown in equations (12), (13).

$$\mu_{A_C}(x)=\begin{cases}1, & if \quad SM(A_C,x)\geq\rho \\ 0, & Otherwise\end{cases} \tag{12}$$

Where $SM(.,\ .)$ is a function for computing the similarity measure between input *x* and fuzzy set A_C, and ρ is a predefined threshold. Likewise, elements that do not belong to concept C will have a unity membership function for the concept $\neg C$ of:

$$\mu_{A_{\neg C}}(x)=\begin{cases}1, & if \quad SM(A_{\neg C},x)\geq\rho \\ 0, & Other\ wise\end{cases} \tag{13}$$

METHODOLOGY

Hierarchical complementary learning method

Learning algorithm stages

The proposed system works in two levels of different accuracy. In the first level there is no need to high accuracy. Therefore the importance of all features will be treated the same. Specifically in the second level features should be classified (categorized) according to their importance to find the output classes. In other words, in the second level each feature has a degree of importance which determines how much that feature is significant to find the output classes.

The following provides the learning steps details:

Step 1: Split instances to positive and negative samples ($D\rightarrow D^+\bigcup D^-$). This step is actually similar to the human's knowledge registration in the brain, where different brain areas are registered for different objects.

Step 2: Remove redundant features (genes) using the previously described Fuzzy-Rough feature reduction.

Step 3: After removing redundant features (genes), for each remaining feature an absolute difference of positive mean value from the negative one as: $diff=|\mu_+-\mu_-|$, will be defined. This metric distance determines how much each feature is important to find the output classes.

As shown in Figure 1(b) the proposed classifier has two subsystems working in two different levels of accuracy. First level is fed with all features and all are treated the same. Second level employs the defined metric to feed features to the subsystems; features with larger metric value are more significant to the outcome and vice versa.

Step 4: The trapezoidal fuzzy sets is defined for each feature (Bezdek, 1981), to produce a rule base for positive and negative samples separately (Chatterjee, 2004), that is $R=R^+\bigcup R^-$, where:

$$R^+=\bigcup_{k=1}^{K^+}r_k\,,R^-=\bigcup_{k=1}^{K^-}r_k\,,\quad k\in[1,k] \tag{14}$$

K is the total number of rules, that is $K=K^++K^-$. Each rule is in the form of:

$$\begin{aligned}&r_k: IF\quad x_1\quad is\quad A_{1k},...,x_i\quad is\quad A_{ik},...,x_A\quad is\quad A_{Ak},\\&THEN\quad y_1\quad is\quad B_{1k},...,y_m\quad is\quad B_{mk},...,y_M\quad is\quad B_{Mk}\end{aligned} \tag{15}$$

In which x_i is the i*th* input attribute, y_m is m-dimension output, A_{ik} is the input linguistic term that links x_i to the k*th* rule, and B_{mk} is the output linguistic term that links the k*th* rule to y_m.

Step 5: by chunking and tuning fuzzy sets and the rule combination interpretability of the system is improved. For more information about this step please refer to our previous work (Shaeiri, 2011).

Step 6: Once discarding redundant rules and the initial tuning of fuzzy sets is over, validity of the rules should be assessed. It requires the definition of the rule fitness which is set to the number of samples that the rule has correctly classified during the training process. The rule is removed if its fitness is lower than a predefined threshold.

Inference method

The strategy of the inference process is as follows: For each sample, which is fed to the classifier, some rules will be activated, we calculate average and maximum of these activated rules. Then the complementary mechanism is exploited for the inference stage. Equations (17) and (18) show the inference process for the first level and the second level respectively:

$$O=\begin{cases}O^+, & \text{if} \quad \mu_{R^+}(x)>\mu_{R^-}(x)\\ O^-, & \text{if} \quad \mu_{R^+}(x)<\mu_{R^-}(x)\end{cases} \tag{16}$$

In which:

$$\mu_{R^+}(x) = \left(\sum_{k=1}^{K^+}\sum_{i=1}^{A}\mu_{ki}(x)\right) > \left(\sum_{k=1}^{K^-}\sum_{i=1}^{A}\mu_{ki}(x)\right) \cap$$
$$\max_{k\in K^+}\left\{\sum_{i=1}^{A}\mu_{ki}(x)\right\} > \max_{k\in K^-}\left\{\sum_{i=1}^{A}\mu_{ki}(x)\right\} \tag{17}$$

$$\mu_{R^-}(x) = \left(\sum_{k=1}^{K^-}\sum_{i=1}^{A}\mu_{ki}(x)\right) > \left(\sum_{k=1}^{K^+}\sum_{i=1}^{A}\mu_{ki}(x)\right) \cap$$
$$\max_{k\in K^-}\left\{\sum_{i=1}^{A}\mu_{ki}(x)\right\} > \max_{k\in K^+}\left\{\sum_{i=1}^{A}\mu_{ki}(x)\right\}$$

If the decisions from the maximum and average do not tally, the system employs the second level to perform the inference process. The inference process of the second level is as the same as the first level; the only difference is that value of the membership degree of each feature in the second level is the sum of the membership degrees of low level features. The inference process of the second level is formulated in equation (18).

$$\mu_{R^+}(x) = \left(\sum_{k=1}^{K^+}\sum_{i=1}^{A}\sum_{j=1}^{i}\mu_{kj}(x)\right) > \left(\sum_{k=1}^{K^-}\sum_{i=1}^{A}\sum_{j=1}^{i}\mu_{kj}(x)\right) \cap$$
$$\max_{k\in K^+}\left\{\sum_{i=1}^{A}\sum_{j=1}^{i}\mu_{kj}(x)\right\} > \max_{k\in K^-}\left\{\sum_{i=1}^{A}\sum_{j=1}^{i}\mu_{kj}(x)\right\} \tag{18}$$

$$\mu_{R^-}(x) = \left(\sum_{k=1}^{K^-}\sum_{i=1}^{A}\sum_{j=1}^{i}\mu_{kj}(x)\right) > \left(\sum_{k=1}^{K^+}\sum_{i=1}^{A}\sum_{j=1}^{i}\mu_{kj}(x)\right) \cap$$
$$\max_{k\in K^-}\left\{\sum_{i=1}^{A}\sum_{j=1}^{i}\mu_{kj}(x)\right\} > \max_{k\in K^+}\left\{\sum_{i=1}^{A}\sum_{j=1}^{i}\mu_{kj}(x)\right\}$$

Employing the complementary learning results two separate systems which are working parallel. Thus the parallel action saves the time of computation and the training to produce an output. In the inference process for some samples with conflict between them, some constraints may arise. In this case majority voting scheme will be used. This means, each positive rule makes a positive decision if the positive degree is more than a predefined threshold; otherwise a negative decision is made. Likewise, for negative rules:

$$Decision^+ = \begin{cases} +, & if \quad \mu_{r_k^+}(x) > \rho \\ -, & otherwise \end{cases} \tag{19}$$

$$Decision^- = \begin{cases} -, & if \quad \mu_{r_k^-}(x) > \rho \\ +, & otherwise \end{cases} \tag{20}$$

The overall decision of the positive and negative rules is taken as the majority votes from the rules within the subsystems, respectively (equations (21), (22)). The final decision is made based on the output of these two subsystems. If both reach the consequence, then a decision is outputted, otherwise a decision is chosen randomly.

$$D^+(x) = \begin{cases} +, & if \quad \left|Decision^+ = +\right| > \left|Decision^- = -\right| \\ -, & otherwise \end{cases} \tag{21}$$

$$D^-(x) = \begin{cases} -, & if \quad \left|Decision^- = +\right| > \left|Decision^+ = -\right| \\ +, & otherwise \end{cases} \tag{22}$$

RESULTS

The datasets

Three public microarray datasets were used to assess the performance of the proposed classifier (gene expression data, 2012). The following is a brief description of these datasets.

1. Leukemia: The Leukemia dataset includes 7,129 genes and 72 patients. It is classified into two types of cancer: Acute Lymphocytic Leukemia (ALL) and Acute Myeloid Leukemia (AML). Among them 47 of the samples were from ALL patients. An additional 25 cases were from patients with AML. Employing the Fuzzy-Rough Set feature selection for this dataset resulted in selecting 4 genes.
2. Prostate: This dataset consists of 102 samples from the same experimental conditions, in two classes of tumor and normal, which have 52 and 50 samples respectively. Each sample is described using 12,600 genes. Using the Fuzzy-Rough Set feature selection for this dataset resulted in selecting 9 genes.
3. Colon: This dataset also contains 62 samples collected from colon cancer patients. Among these samples, 40 samples are tumor type and 22 are of the normal type. There are 2,000 genes selected based on the confidence in the measured expression levels. Gaining the Fuzzy-Rough Set feature selection for this dataset resulted in selecting 9 genes.

For evaluating the performance, leave-one-out-cross-validation (LOOCV), is applied. In the ten-fold cross validation approach samples are divided in ten disjoint subsets of equal size. Samples are selected on basis of nine of these subsets, and then the remaining subset (the validation subset), is used to estimate the predictive accuracy of the trained classifier, using only the selected samples.

This process is repeated ten times, each time leaving one set out for testing and the others for training. The

Table 1. Accuracy comparison of Classifiers in the three benchmarked datasets.

Method / Dataset	Leukemia	Prostate	Colon	Average
PLR	98.20	96.37	99.35	97.97
HCL	82.44	58.47	65.22	68.71
SVM-RFE-linear SVM	96.65	94.10	93.66	94.80
JCFO	100	89.80	96.80	95.53
RVM	97.20	91.34	88.14	92.22
PCA-SVM	66.70	87.00	65.40	73.03
ISO-SVM	91.76	95.77	100	95.84
PCA-C4.5	70.84	100	65.4	78.75
ULDA	97.67	92.04	82.24	90.65
LLE-C4.5	96.00	100	95.01	97.00
Proposed method	100	98.06	98.57	98.87

cross-validation accuracy rate is given by the average of the ten estimates of the predictive accuracy rate.

To demonstrate the significance of the proposed method, it is compared with some recently published methods: Penalized Logistic Regression (Shen, 2005), SVM-RFE Feature Selection and Linear SVM Classifier (Duan, 2005), JCFO, RVM (Krishnapuram, 2004), PCA-SVM, LDA-SVM, ISO-SVM, PCA-C4.5, ULDA (Ye, 2004) and LLE-C4.5 (Lee, 2008). Table 1 contains the accuracy comparison of the proposed method with respect to these recent methods in the three benchmarked datasets.

The proposed method is more comparable to HCL system (Tan, 2008) which has a hierarchical complementary structure, But some limitations can be observed in HCL. In HCL augmented variance ration (AVR) is used as the feature ranking preprocessing step. This drawback can be overcome by using an adaptive feature selection method. Another drawback is that its knowledge is established based on single feature instead of features combination. These two limitations are solved in the proposed method thanks to using Fuzzy-Rough feature selection preprocessing step and the inference mechanism of the classifier. It can be seen from Table 1 that the proposed method is superior than the recently methods.

DISCUSSION

One of the significant strengths of the proposed method is using the Fuzzy-Rough Set feature selection as a preprocessing stage. Fuzzy-Rough Set is a powerful tool for feature reduction in real-valued and noisy (or mixture of both) datasets. Therefore, it reduces the number of features in the gene expression dataset (a real-valued and very noisy dataset), effectively without the need for user-supplied information such as thresholds or any extra expert defined parameters. The only additional information required is in the form of fuzzy sets which can be automatically derived from the dataset. Furthermore, Fuzzy-Rough Set feature selection preserves the semantics of surviving features after removing any redundant ones. This is an important aspect in satisfying the requirement of user readability of the generated knowledge model, as well as ensuring the understandability of the pattern classification process.

As a result of employing Fuzzy-Rough Set for the feature selection in the present work, robustness against the noise and retaining the semantics of the datasets is achieved.

Using the complementary learning method needs splitting the dataset into two subsets of positive and negative samples that should be processed simultaneously (in parallel). This also results in reducing the complexity and time consuming of the system. Using a self-organizing rule creation, proposed method takes into account the minimum and maximum value of each feature. In the training process the kernel of fuzzy sets expands, whilst preventing the overlap between fuzzy sets of opposite classes. This feature in favor of the complementary learning mechanism, results in alleviating effect of outliers as well as good interpretability. A conflict arising is more likely in the gene expression datasets, which is avoided, thanks to using the two level structure system, in which each level has different degree of accuracy.

Conclusion

In this paper a new method for classification of gene expression data is proposed. The proposed method has a preprocessing stage, in which the number of features is reduced using the Fuzzy-Rough Set feature selection.

Only those genes that play a pivotal role in discerning between output classes are retained. Then, a new fuzzy 2-level complementary system for the classification is proposed. As Table 1 shows, the proposed method is very competitive in terms of accuracy with respect to the other methods. High accuracy in the test experiment, as a sign of good generalization property of the system, is achieved. Since the proposed system exploits the complementary learning and the hierarchical organization (as some aspects of human cognitive ability), a complexity reduction and good interpretability are achieved. Furthermore, the Fuzzy-Rough Set is found a powerful mathematical tool in selecting relevant features of real-valued datasets. It is a semantic preserving dimensionality reduction technique (feature selection) and encapsulates two complementary types of uncertainty at once. Thus, employing Fuzzy-Rough Set feature selection on the dataset a minimal subset of features with sufficiently high accuracy in representing the original data is obtained.

REFERENCES

Bezdek JC (1981). Pattern recognition with fuzzy objective function algorithms. New York. Plenum Press.

Cai J, Dayanik A, Yu H, Hasan N, Terauchi T, Grundy W (2000). Classification of gene cancer types by support vector machines using microarray gene expression data. Int. Conf. Intell. Sys. Mol. Bio.

Chatterjee A, Rakshit A (2004). Influential rule search scheme: A new fuzzy pattern classifier. IEEE. Trans. knowl. data eng., 16: 881-893.

Dash M, Liu H (1997). Feature selection for classification. Intell. Data Anal., 1: 131-156.

DeJong KA, Spears WM, Gordon DF (1993). Using genetic algorithms for concept learning. Mach. Lear. 16: 161-188.

Devijver P, Kittler J (1982). Pattern recognition: A statistical approach. Prent. Hall.

Ding C (2002). Analysis of gene expression profiles: Class discovery and leaf ordering. RECOMB. pp. 127-136.

Duan KB, Rajapakse JC, Wang H, Azuaje F (2005). Multiple SVM-RFE for gene selection in cancer classification with expression data. IEEE. Trans. Nanobiosci., 4: 228-234.

Dudoit S, Fridlyand J, Speed T (2000). Comparison of discrimination methods for the classification of tumors using gene expression Data. Berekly. University of California.

Gauthier I, Skudlarski P, Gore JC, Anderson AW (2000). Expertise for cars and birds recruits brain areas involved in face recognition. Nature. Neuro., 3: 191-197.

Guh L, Song Q, Kasabov N (2004). A novel feature selection method to improve classification of gene expression data. 2nd Asia. Pacif. Bioinf. Conf., 161-166.

Guyon I, Weston J, Barnhill S, Vapnik V (2002). Gene selection for cancer classification using support vector machines. Mach. Learn., 46: 389-422.

Han J (2001). Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nature. Med., 7: 673-679.

Jensen R, Shen Q (2001). A rough set-aided system for sorting www bookmarks. Web Intell. Res. Dev. 95-105.

Jensen R, Shen Q (2004). Selecting informative features with fuzzy-rough sets and its application for complex systems monitoring. Pat. Rec., 37: 1351-1363.

Jensen R, Shen Q (2007). Fuzzy-rough sets assisted attribute selection. IEEE. Trans. Fuz. Sys. 15: 73-89.

Joo MG, Lee JS (2005). A class of hierarchical fuzzy system with constraints on the fuzzy rules. IEEE. Trans. Fuz. Syst., 13: 194-203.

Kira K, Rendell LA (1992). The feature selection problem: Traditional methods and a new algorithm. 9th. Nat. Conf. Artif. Intell., 129-134.

Kohane IS, Kho AT, Butte AJ (2003). Microarrays for an integrative genomics. London. MIT Press.

Kohavi R, John GH (1997). Wrapper for feature subset selection. Artif. Intell., 97: 273-324.

Krishnapuram B, Hartemink AJ, Carin L, Figueiredo M (2004). A Bayesian approach to joint feature selection and classifier design. IEEE. Trans. Pat. Anal. Mach. Lear., 26: 1105-1111.

Langley P (1994). Selection of relevant features in machine learning. AAAI. Fall. Symp., 140-144.

Lee G, Rodriguez C, Madabhushi A (2008). Investigating the efficacy of nonlinear dimensionality reduction schemes in classifying gene and protein expression studies. IEEE. ACM. Trans. Comput. Biol. Bioinf., 5: 368-384.

Ohno-Machado L, Vinterbo S, Weber G (2002). Classification of gene expression data using fuzzy logic. J. Intell. Fuzz. Syst., 12: 19-24.

Pawlak Z (1982). Rough sets. Int. J. Inf. Comp. Sci., 11: 341-356.

Shaeiri Z, Ghaderi R (2011). Genetic diagnosis of cancer by fuzzy-rough gene selection and complementary hierarchical fuzzy classhfier. Bio-Med. Mat. Eng., 21: 37-52.

Shen L, Chong E (2005). Dimension reduction-based penalized logistic regression for cancer classification using microarray data. IEEE. ACM. Trans. Comp. Biol. Bioinf., 2: 166-175.

Tan TZ, NG GS, Quek C (2008). A novel biologically and psychologically inspired fuzzy decision support system: hierarchical complementary learning. IEEE. ACM. Trans. Comp. Biol. Bioinf., 5: 67-77.

Xing WL, Cheng J (2006). Frontiers in biochip technology. New york. Springer.

Xiong M, Li W, Zhao J, Jin L, Boerwinkle E (2001). Feature (gene) selection in gene expression-based tumor classification. Molecul. Gen. Metabo., 73: 239-247.

Ye J, Li T, Xiong T, Janardan R (2004). Using uncorrelated discriminant analysis for tissue classification with gene expression data. IEEE. ACM. Trans. Comp. Biol. Bioinf., 1: 181-190.

Some Gene Expression datasets are available online (2012). http://www.ailab.si/supp/bi-cancer/projections/index.htm.