Journal of Particle Science and Technology 3 (2017) 175-186

J | P | S | T

http://jpst.irost.ir

Journal of Particle Science and Technology

IROST

Diagnosis of the disease using an ant colony gene selection method based on information gain ratio using fuzzy rough sets

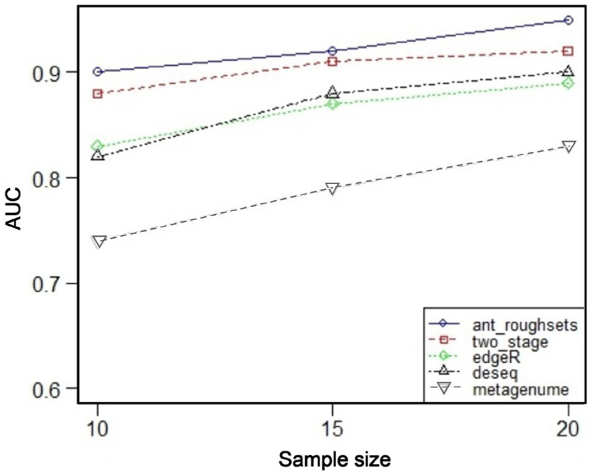
Mohammad Masoud Javidi*, **Sedighe Mansoury**

Faculty of Mathematics and Computer, Department of Computer Science, Shahid Bahonar University of Kerman, Kerman, Iran

HIGHLIGHTS

- Gene selection as a preprocessing phase is very important in the diagnosis of diseases.
- By applying a two-stage gene selection method, the accuracy of detecting diseases process was increased.
- By detecting the genes which were statistically differentially abundant in different phenotypes, the genes that related to healthy or diseases were detected.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 11 December 2017
Revised 17 January 2018
Accepted 24 January 2018

Keywords:
Microarray
Gene selection
Phenotype
Fuzzy rough set
Ant colony optimization
Gain ratio

ABSTRACT

With the advancement of metagenome data mining science has become focused on microarrays. Microarrays are datasets with a large number of genes that are usually irrelevant to the output class; hence, the process of gene selection or feature selection is essential. So, it follows that you can remove redundant genes and increase the speed and accuracy of classification. After applying the gene selection, the dataset is reduced and detection of differentially abundant genes facilitated with more accuracy. This will, in turn, increases the power of genes which are correctly detected statistically differentially abundant in two or more phenotypes. The method presented in this study is a two-stage method for functional analysis of metagenomes. The first stage uses a combination of the filter and wrapper gene selection method, which includes the ant colony algorithm and utilizes fuzzy rough sets to calculate the information gain ratio as an evaluation measure in the ant colony algorithm. The set of features from the first stage is used as input in the second stage, and then the negative binomial distribution is used to detect genes which are statistically differentially abundant in two or more phenotypes. Applying the proposed method on a microarray dataset it becomes clear that the proposed method increases the accuracy of the classifier and selects a subset of genes that have a minimum length and maximum accuracy.

* *Corresponding author: Tel.: +9834-31322251 ; Fax: +9834-33257159 ; E-mail address: javidi@uk.ac.ir*

DOI: 10.22104/jpst.2018.2649.1102

1. Introduction

In the last two decades, the advent of the DNA microarray data set has stimulated a new movement of research into bioinformatics and machine learning. All cells have a nucleus and inside the nucleus, there is DNA. DNA has coding and encoding sections, the coding sections are known as genes. The genes in each individual have different abundances known as gene expression. Each gene performs essential work in any organism [1]. Advances in molecular genetic technologies, such as the micro-arrays of DNA, allow us to obtain a general view of the cell and we can observe expression of a large number of genes [2]. The general process of obtaining gene expression data from a DNA microarray is presented in Figure 1, where the dataset is formed for two classes of normal and diseased. In order to detect differentially abundant genes for different classes and to study their effects on diseases we need to analysis the gene expression dataset. The large dimensions of the dataset lead to statistical and analytical problems, and also, there are very small samples compared to the number of genes in the dataset. In addition, the presence of noise in the genes makes it difficult to detect the specific genes that cause the disease. The application of gene selection is a good approach to overcome to these problems. Using gene selection redundant and irrelevant genes are deleted; thereby, reducing processing time and also diminishing the interference of noisy or unwanted information leading to incorrect classification, in other words the accuracy of the classifier is increased [3]. Gene selection methods can be divided into four categories: filter, wrapper, embedded and hybrid, which are shown in Table 1 [2]. After applying a gene selection method, diseases or tumors is done are determined by detecting differentially abundant genes in two or more phenotypes. Applying an appropriate method to correctly detect these genes is essential. Statistical procedures play a critical role in detecting differentially abundant genes. In this paper, a two stage method is proposed to determine whether a person is ill or healthy. In the first phase, the dimension of the dataset is reduced by applying an ant colony gene selection evaluation method based on information gain ratio that is calculated by fuzzy rough sets. Then in the second phase, a negative binomial distribution is used to determine the health or sickness. The proposed method can be applied to the comparison of more than two microbial conditions; two microbial conditions; so, our method can be applicable to more general situations.

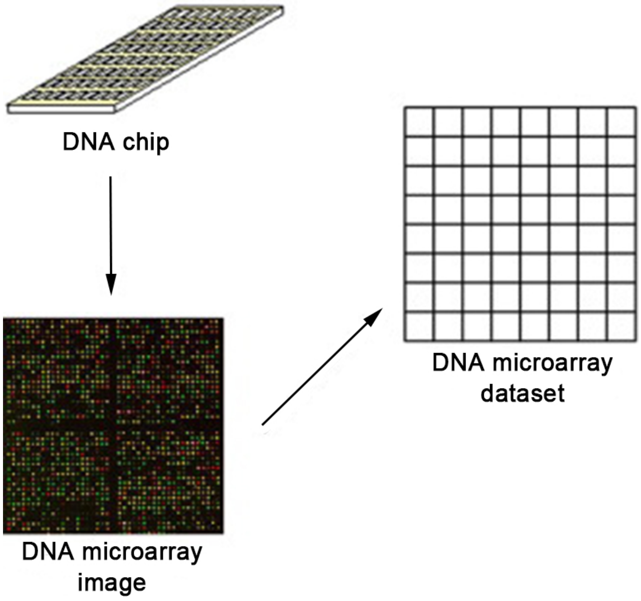


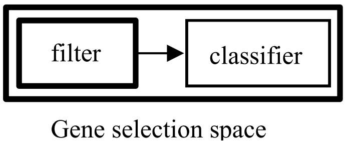
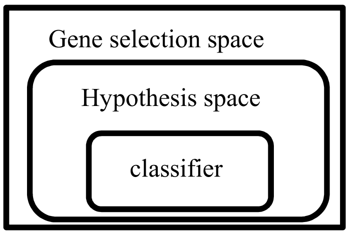
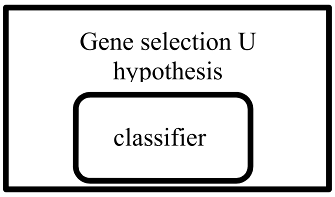
Fig. 1. General process of acquiring gene expression data from the DNA microarray.

2. Related works

In recent years, gene selection has received much attention. Many optimization algorithms of feature or gene selection have been presented to increase classification accuracy. The concept of gene selection is viewed as one of the most important techniques in Rough Set Theory. There are many feature selection methods that use rough sets. Inbarani et al. [4] presented a supervised hybrid feature selection algorithm based on particle swarm optimization (PSO) and rough sets. This method applies a positive region-based dependency measure to calculate the dependency of the decision feature on the conditional features, which is suitable only for smaller datasets. Chen et al. [5] present a rough set-based feature selection method using the Fish Swarm Algorithm. This algorithm uses a rough set-based dependency measure and thus is not suitable for large datasets. Park and Choi [6] proposed information-theoretic dependency roughness. This algorithm considers the information-theoretic attribute dependency degrees of categorical-valued information systems. The execution time of this method is not provided.

The majority of studies on rough sets have been focused on constructive approaches. In the Pawlak's

Table 1. Comparison of General Schemes for Gene Selection Methods [2].

Method	Advantages	Disadvantages
Filter method  	➢ Easily scaled to very high-dimensional datasets. ➢ Very fast and are computationally simple. ➢ Not dependent on any particular algorithm. ➢ Feature selection is to be carried out only once, and then different classifiers can be evaluated. ➢ Time complexity is O(n), which is low as Compared to other methods Simple.	➢ Do not take into account the interaction with the classifier. ➢ Each feature is measured separately and thus does not take into account the feature dependencies. ➢ Lack of feature dependencies results in the degraded performance as compared to other techniques. ➢ Creates redundancy. ➢ Evaluates genes based on their individual scores ignores their relevance in combination with other genes.
Wrapper method  	➢ Involve the interaction between model selection and feature subset search. ➢ Take feature dependencies into account. ➢ Implementing a wrapper method is quite easy and straightforward in supervised learning. ➢ Tests the predictive power of genes. ➢ Carries out exhaustive search, generating. ➢ optimal solutions.	➢ These methods have to overfit with a higher risk than filter techniques. ➢ Wrapper methods are computationally intensive. ➢ Exponential time complexity. ➢ Doesn't take enough measures to eliminate redundancy.
Embedded method  	➢ Interacts with the classifier. ➢ Achieves computational complexity better than wrapper methods. ➢ Models feature dependencies. ➢ Tests the predictive power of genes fitting.	➢ Classifier dependent selection. ➢ Prone to over-fitting.
Hybrid method	➢ Can combine the advantages of various approaches.	➢ Time complexity may increase.

rough set model [7], the correlation relationship is a key concept. However, this correlation relationship is a very stringent condition that can restrict the application domain of the rough set model. To solve this problem a fuzzy similarity relation can be used to replace an equivalence relation, which was called the fuzzy rough set. Applying fuzzy rough sets in gene selection has received much attention. Gene or feature selection by fuzzy rough sets was first proposed by Wang et al. [8]. They evaluated the hypoxic resistance of a patient on the basis of the values of his blood pressure during a barocamera examination. The measurements were evaluated by the FRS criteria. Jenson and Shen [9] proposed a feature selection method which uses the dependency function to compute the importance of attributes by fuzzy rough sets. Pradipta and Partha [10] proposed a feature selection where the fuzzy rough set was used to measure the relevance and significance of features. In [11] a method is presented that uses consistence degree as a critical value to reduce redundant attributes in a database. In this approach, a rule based classifier applying a generalized fuzzy-rough set model is proposed. This classifier is effective on noisy data. In [12] a feature selection method with fuzzy-rough and ant colony optimization, similar to our method, is provided. However, the entropy value is used in this method. The disadvantage of this method is that the optimal subset may not be properly selected, because in some cases the entropy criterion (a gene with many distinct values) is high causes the algorithm to selects this gene although it may not be the proper gene. In this paper we presented a method that applies the information gain ratio criteria as the evaluation measure, so it can select the proper genes.

Several statistical methods have been developed to compare various microbial communities in terms of detecting differentially abundant genes, e.g., SONs [13], XIPE-TOTEC [14], Metastates [15] and MEGAN [16]. However, these methods are designed to compare exactly two phenotypes. The Shot Gun Functionalize R [17] method is based on regression and is useful in data with more than two phenotypes; however, the disadvantage of this method is that it only works with discrete data that has a Poisson distribution. Poisson distribution is not flexible for discrete data that has high dispersion. In this paper, we proposed a hybrid gene selection method that uses ant colony and fuzzy rough sets in order to calculate information gain ratio as an evaluation criteria. After selecting an optimal subset of genes, this subset is used in negative binomial (NB) distribution. The NB distribution is widely used to model count data.

3. Some basic notations

In this section, we briefly describe the theory of rough set and also information measures in rough and fuzzy-rough sets theory. Rough set theory was proposed by Pawlak [7]. The concept of a rough set has been proposed as a new mathematical tool to deal with uncertain and imprecise data. This theory has been accepted from the beginning, and has been used in many fields of data analysis such as banking [18], economics and finance [19], medical imaging [20], medical diagnosis [21], and data mining [22].

3.1. Basic rough set notation

Let, $IS=<U, A, V, f>$, be an information system, where U is a nonempty set of finite object, A is a finite set of attributes or genes, and V is the union of attribute domains, where V_a is the set of values for the attribute a; f: $A\times U\rightarrow V$ is an information function that appropriate special values from the domains of attribute to object. If $P\subseteq A$, then an associated indiscernibility equivalence relation, $IND(P)$, is defined as [23]:

$$IND(P)=\{(x,y)\in U^2\,\forall \mid a\in P f(a,x)=f(a,y)\} \tag{1}$$

Since $IND(P)$ is a reflexive, symmetric, and transitive relation, it is an equivalence relation; therefore, $IND(P)$ can create a partition on U that is denoted by $U|IND(P)$ or more simply $U|P$, and $[X]_P$ represents an equivalence class of $IND(P)$ containing x. The lower and upper estimates for $X\subset U$, respectively, are defined as follows [23]:

$$P\downarrow X=\{x\in U \mid [x]_P\subseteq X\} \tag{2}$$

$$P\uparrow X=\{x\in U \mid [x]_P\cap X\neq\varnothing\} \tag{3}$$

Based on the lower and upper estimates, the boundary regain is defined as follows [23]:

$$BND_p(X)=P\uparrow X-P\downarrow X \tag{4}$$

3.2. Information measures in rough set theory

Assume $X_i\in U|IND(P)$ and $X_j\in U|IND(Q)$ are partitions of U which are induced by P and Q, respectively. The probability distribution of X_i is defined as follows and the probability distribution of X_iX_j is defined as Eq. (6), where |..| denotes the cardinality [23].

$$P(X_i)=\frac{|X_i|}{|U|} \tag{5}$$

$$P(X_iX_j)=\frac{|X_i\cap X_j|}{|U|} \tag{6}$$

Definition 1: If $IS=<U, A, V, f>$ is an information system, B is a subset of A and $X_i\in U|B$, then the Shannon's entropy $H(B)$ of B is defined as [23]:

$$H(B)=-\sum_{i=1}^{n}P(X_i)\log P(X_i)=-\sum_{i=1}^{n}\frac{|X_i|}{|U|}\log\frac{|X_i|}{|U|} \tag{7}$$

Definition 2: In information system $IS=<U, A, V, f>$, the join entropy of P and Q is defined as [23]:

$$\begin{aligned}H(PQ)=H(P\cup Q)&=-\sum_{i=1}^{n}\sum_{j=1}^{m}P(X_iX_j)\log P(X_iX_j)\\&=-\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{|X_i\cap X_j|}{|U|}\log\frac{|X_i\cap X_j|}{|U|}\end{aligned} \tag{8}$$

where $X_i\in U|B$, $X_j\in U|Q$, and $P,Q\subseteq A$.

Definition 3: The conditional entropy of D with condition B for decision system $DS=<U,C\cup D,V,f>$ is defined as [23]:

$$(D\,|\,B)=-\sum_{i=1}^{n}P(X_i)\sum_{j=1}^{m}P(X_j\,|\,X_i)\log P(X_j\,|\,X_i) \tag{9}$$
$$=-\sum_{i=1}^{n}\frac{|X_i|}{|U|}\sum_{j=1}^{m}\frac{|X_i\cap X_j|}{|U|}\log\frac{|X_i\cap X_j|}{|U|}$$
$$=-\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{|X_i\cap X_j|}{|U|}\log\frac{|X_i\cap X_j|}{|U|}$$

where, B is a subset of C, and C is the condition attribute set; $X_i \in U|B$ and $X_j \in U|D$, where D is the decision attribute.

Definition 4: The mutual information of B and D is defined as follows [23]:

$$I(B;D)=H(D)-H(D\,|\,B) \tag{10}$$

Definition 5: The gain of attribute $a \in C\text{-}B$ is defined as [23]:

$$\begin{aligned}Gain(a,B,D)&=I(B\cup\{a\};D)-I(B;D)\\&=H(D\,|\,B)-H(D\,|\,B\cup\{a\})\end{aligned} \tag{11}$$

Definition 6: The mutual information gain ratio of attribute a, is defined as [23]:

$$\begin{aligned}Gain\_Ratio(a,B,D)&=\frac{Gain(a,B,D)}{H(\{a\})}\\&=\frac{I(B\cup\{a\};D)-I(B;D)}{H(\{a\})}\end{aligned} \tag{12}$$

3.3. Information measures in fuzzy-rough set theory

In fuzzy rough sets, it is essential to define a fuzzy equivalence relation. $\tilde{R}$ is a fuzzy equivalence relation, if it satisfies:

Reflectivity: $\tilde{R}(x,y)=1,\forall x\in X$

Symmetry: $\tilde{R}(x,y)=\tilde{R}(y,x),\forall x,y\in X$

Transitivity: $\tilde{R}(x,y)\geq\min\limits_{y}\{\tilde{R}(x,y),\tilde{R}(y,z)\}$

$M(\tilde{R})$ represents a relation matrix for x_i, $x_j \in X$, that $\tilde{R}$ is a fuzzy equivalence relation defined on a nonempty finite set X.

$$M(\tilde{R})=\begin{pmatrix} r_{11} & \cdots & r_{1n}\\ \vdots & \ddots & \vdots\\ r_{n1} & \cdots & r_{nn}\end{pmatrix} \tag{13}$$

Here, $r_{ij}\in[0,1]$ is the relation value of x_i and x_j that can be written as $\tilde{R}(x,y)$. For the crisp rough set model, if x_i equals to x_j with respect to the crisp equivalence relation R then r_{ij}= 1; otherwise, r_{ij}= 0. A similarity function that has been used to calculate the equivalence relation is shown by Eq (14), where x_i and x_j are attribute values of two objects on attribute a; a_{max} and a_{min} are maximal and minimal values of attribute a, respectively [23].

$$r_{ij}=\begin{cases}1-4\times\dfrac{|x_i-x_j|}{|a_{\max}-a_{\min}|}, & \dfrac{|x_i-x_j|}{|a_{\max}-a_{\min}|}\leq 0.25\\ 0\end{cases} \tag{14}$$

Two important operators in the fuzzy equivalence relation that are useful for implementing fuzzy theory are defined by [23]:

$$\tilde{R}=\tilde{R}_1\cup\tilde{R}_2\Leftrightarrow\tilde{R}(x,y)=\max\left\{\tilde{R}_1(x,y),\tilde{R}_2(x,y)\right\}$$

$$\tilde{R}=\tilde{R}_1\cap\tilde{R}_2\Leftrightarrow\tilde{R}(x,y)=\min\left\{\tilde{R}_1(x,y),\tilde{R}_2(x,y)\right\}$$

Definition 7: The fuzzy partition of the universe U generated by $\tilde{R}$, is defined as [23]:

$$U/\tilde{R}=\left\{[x_i]_{\tilde{R}}\right\}_{i=1}^{n} \tag{15}$$

Here, $\tilde{R}$ is a fuzzy equivalence relation and $[x]_{\tilde{R}}$ is the fuzzy equivalence class equal to $\frac{r_{i1}}{x_1}+\frac{r_{i2}}{x_2}+\ldots+\frac{r_{in}}{x_n}$.

Definition 8: The cardinality $[x]_{\tilde{R}}$ is defined as [23]:

$$|[x_i]_{\tilde{R}}|=\sum_{j=1}^{n}r_{ij} \tag{16}$$

Definition 9: Information quantity of the fuzzy attribute set or the fuzzy equivalence relation is defined as [23]:

$$H(\tilde{R})=-\frac{1}{n}\sum_{i=1}^{n}\log\frac{|[x_i]_{\tilde{R}}|}{n} \tag{17}$$

Definition 10: The joint entropy of B and E is defined as [23]:

$$H(BE)=H(\tilde{R}_B\tilde{R}_E)=-\frac{1}{n}\sum_{i=1}^{n}\log\frac{|[x_i]_{\tilde{B}}\cap[x_i]_{\tilde{E}}|}{n} \tag{18}$$

where FIS= <U, A, V, f> is a fuzzy information system, A is the attribute set, and B and E are two subsets of A.

Definition 11: Let FIS= <U, A, V, f> is a fuzzy decision system, C is the condition attribute set, D is the decision attribute and $B \subseteq C$. The condition entropy D on condition B can be calculated as follows [23]:

$$\tilde{H}(D \mid B) = -\frac{1}{n}\sum_{i=1}^{n}\log\frac{|[x_i]_{\tilde{B}} \cap [x_i]_{\tilde{D}}|}{|[x_i]_{\tilde{B}}|} \tag{19}$$

In the above relation, $[x_i]_{\tilde{B}}$ and $[x_i]_D$ are fuzzy equivalence classes containing x_i generated by B and D, respectively.

Definition 12: The mutual information of B and D is defined as [23]:

$$\tilde{I}(B;D) = \tilde{H}(D) + \tilde{H}(B) - \tilde{H}(BD) \tag{20}$$

Definition 13: In decision system FDS= $<U,C\cup D,V,f>$, $\forall a \in C\text{-}B$ the gain of attribute a, can be defined as [23]:

$$Gain(a,B,D) = \tilde{I}(B \cup \{a\};D) - \tilde{I}(B;D) \tag{21}$$

Definition 14: According to the definition of 13, the mutual information gain ratio of attribute a, can be defined as [23]:

$$GainRatio(a,B,D) = \frac{Gain(a,B,D)}{\tilde{H}(\{a\})} = \frac{\tilde{I}(B \cup \{a\};D) - \tilde{I}(B;D)}{\tilde{H}(\{a\})} \tag{22}$$

4. Proposed method

4.1. Gene selection phase

In this section, a new filter-wrapper approach for gene selection in fuzzy-rough sets is described. In this approach, filter phase employs a modified ACO search strategy which is able to do gene selection function as a multi-modal problem, and the wrapper phase includes a learning model that evaluates the chosen subsets of genes from the filter phase and select the best subset, then calculates pheromones changes in the selected subsets. Choosing the subsets of features with first and second maximum accuracies as candidate subsets for minimal data reductions is a contribution of this work; so each chosen minimal subset has a short length along with an acceptable accuracy value; consequently, the approach is able to satisfy both an increase the accuracy and a decrease in the length of reduced subsets, concurrently. In detail, in order to implement this approach we need the feature selection problem space to be considered in the form of a complete non-directed graph. The nodes, indicating the genes and edges, represent the probability of choosing the next node. The algorithm starts with the production of k number of ants, which is half the number of genes. The following steps are followed to complete each ant's tour:

1- Initialize ants with random and different nodes.
2- For each ant k, consider set S_K includes all the nodes without initial node, as accessible locations.
3- The ant k chooses the next node according to the transition rule that will be dealt with in the next section.
4- The selected node is removed from the S_K.
5- For each ant k, the third and fourth stage is repeated until S_K is empty.
6- The best answer achieved is saved.

After each ant completes its tour, the pheromone is updated on the routes traversed from origin to destination, according to the algorithm explained in section 3-1-2. At the end of each iteration, the best observed solutions are kept; i.e. in each iteration, we consider the subsets of the genes that have maximum accuracy as the best candidate subsets. We preserve the subsets which have the first and the second maximum accuracies among all the best candidate subsets from the first iteration to the current. Then, we consider the minimal subsets from the preserved subsets as the best of all the iterations. Because the wrapper method utilizes a learning model, gene selection based on wrappers boosts the accuracy of the model; however, this method increases the order of mathematical complexity. In this method, instead of evaluating the genes separately, the subsets found by the filter are evaluated using the wrapper model to decrease the complexity. Output of wrapper model (accuracy of the classifier) is a criterion for the goodness evaluation of the subsets found. After the end of each run the best seen solution, from the first iteration until the current one is saved as an optimal solution. In addition to detecting high quality subsets of genes, finding more than one solution in one run is another advantage of this method compared to other methods.

4.1.1. Transition rule and gene deletion

The transition rule introduced in [24] is used for exploring the nodes' space. Node j, as a candidate for selection, is selected with probability 0.5 using the

$$p_{ij}^{k}=\begin{cases}1, \; j=\arg\max\left\{\tau_{ij}^{\alpha}\eta_{ij}^{\beta}\right\} \\ 0, \quad \mathit{otherwise}\end{cases} \tag{23}$$

If an ant selects a new node, that node is removed from the set of available nodes, and if candidate node j is not selected that candidate node will also be removed from the set of available nodes. In this case, the following relation in the roulette wheel mechanism, as the probability of selecting the available nodes, is used to select the next node.

$$p_{ij}^{k}(t)=\begin{cases}\dfrac{[\tau_{ij}]^{\alpha}.[\eta_{j}]^{\beta}}{\sum_{x\in S_K}[\tau_{ix}]^{\alpha}.[\eta_{x}]^{\beta}}, & j\in s_{k} \\ 0, & \mathit{otherwise}\end{cases} \tag{24}$$

In both of the above equations, α = 0.5 and β = 1, and the initial value of τ_j is equal to 0.1. By selecting each node, in the roulette wheel mechanism, that node and all nodes before it, η_j = *GainRatio* (*j*, N_K, *D*) are calculated by Eq. (22) as heuristic information and N_K is regarded as a set of selected nodes by ant k, and τ_{ij} is the pheromone value of edge ij.

4.1.2. Pheromones updating rules

After each individual ant created its own complete tour, the pheromone is updated on the path it travelled from the beginning to the end, as follows:

1- On each edge of the complete graph, the pheromone evaporates according to equation (25).

2- In each iteration, the pheromone on the path is updated according to equations (26) and (27).

3- In order to maintain the best answers, the pheromone on the best path in all of the repetitions is updated according to (28).

$$\tau^{new}=(1-\rho).\tau^{old} \tag{25}$$

$$\Delta\tau_{ij}=\frac{\gamma'_{N_k}}{lenght(N_k)} \tag{26}$$

$$\tau_{ij}^{new}=\begin{cases}\tau_{ij}^{old}+\Delta\tau_{ij}, & if \quad ij\in BF \\ \tau_{ij}^{old}+\varphi*\Delta\tau_{ij}, & otherwise\end{cases} \tag{27}$$

$$\tau_{ij}^{new}=\tau_{ij}^{old}+\varphi*\gamma'_{N_k} \tag{28}$$

where φ = 0.5, ρ = 0.2 and BF is the best path traversed in the current iteration. γ_{N_k} is the accuracy of the classifier as output of the learning model.

4.2. Differentially abundant feature detection stage

At this stage, the genes which are statistically differentially abundant in two or more conditions are detected. In real metagenomics count data, the variance is usually greater than the corresponding mean of the gene abundance. Negative binomial distribution (NB) is often used for high-dispersion data.

4.2.1. NB model

Suppose r of p genes are selected from the first stage. Let Y be the vector of the numbers of reads for gene i in all samples where i=1, 2, 3,…, r. Each element (y_S) in the Y vector with a negative binomial distribution is modeled as follows:

$$f_Y(y_s;\mu_s,\theta)=\frac{\Gamma(y_s+\theta)}{\Gamma(\theta).y_s!}.\frac{\mu_s^{y_s}.\theta^{\theta}}{(\mu_s+\theta)^{y_s+\theta}} \tag{29}$$

$E(y_S)$=μ_S is the mean and the variance is var(y_S)= $\mu_S(1+\mu_S/\theta)$. The variance is quadratic in the mean. The negative binomial distribution model can also be modeled with the dispersion parameter, ϕ=1/θ. In this case, the mean is equal to and the variance is $\mu_S(1+\phi\mu_S)$. Initially, ϕ is greater than zero, and when $\phi\rightarrow0$, the negative binomial distribution is reduced to the standard Poisson distribution with the parameter μ_S. In the generalized linear model, the logarithmic link is the most appropriate method for linking the mean response μ in negative binomial distribution variable to a linear combiner of predictors *x*. For each gene i (i= 1, 2, 3, …, r), $\log(\mu_s)=x_S^T\beta$ where x_S^T is 1×*K*, the line vector contains the indicative variables of the phenotypes, S=1, 2, …, N, K represents the number of phenotypes and β is the corresponding *K*×1 column vector of unknown regression parameters. Auxiliary variables can be introduced into a regression model based on the NB distribution via the relationship:

$$\log(\mu_s)=\sum_{j=1}^{K}x_{sj}\beta_{j-1} \tag{30}$$

In the negative binomial distribution model for mean $\mu_S=\exp(x_S^T\beta)$, β and ϕ are estimated by maximizing the log likelihood function:

$$l(\beta,\phi;Y)=\sum_{s=1}^{N}\{\log\left(\frac{\Gamma(y_s+\phi^{-1})}{\Gamma(\phi^{-1})}\right)-\log(y_s!) \tag{31}$$

$$-(y_s+\phi^{-1})\log(1+\phi\mu_s)+y_s\log\phi+y_s x_s^T\beta\}$$

5. Results

To implement the proposed method, we use a data set with 20 samples and 1000 genes. This data set has two classes; one is healthy and the other expresses the sickness of the samples. Among these 10 samples are healthy and the rest are patient. In order to implement the proposed method, we utilize the R statistical software on a five-core computer that has 1 gigabyte of RAM. The proposed method has been implemented with a number of different samples of the data set and the results of these experiments were compared with four current reliable methods: Two-stage, edgeR, metagenomeSeq and DESeq. The results are expressed in terms of time, accuracy, ROC, AUC, PR Curve FDR, and the power in detection of the true differentially abundant genes. In addition, a criterion named, Ψ_B as described below, is also examined in the results.

$$\Psi_B = \frac{accuracy(B)}{length(B)} \tag{32}$$

where B is a subset of genes. By increasing the classification accuracy and reducing the length of the selected subset, Ψ_B increases. This indicates an increase in the efficiency of the method, because the efficiency of the feature or gene selection is based on the accuracy and number of selected genes.

In Figure 2, the computational time of the proposed method for sample sizes of 10, 15, and 20, is compared with the other methods. According to Figure 2, the Runtime of the proposed method is less than DESeq and Metastates but longer than the rest of the methods. Therefore, the proposed method is not efficient in terms of run time. The high execution time of this algorithm is due to the implementation of the ant colony algorithm. Microarray data requires several nesting loops with a high repetition rate and the number of high repetitions is due to the fact that microarray data usually has a high number of genes. Another reason for the high execution time of this algorithm is the need for high-dimensional square matrices to calculate entropy and the information gain ratio, which are time consuming calculations.

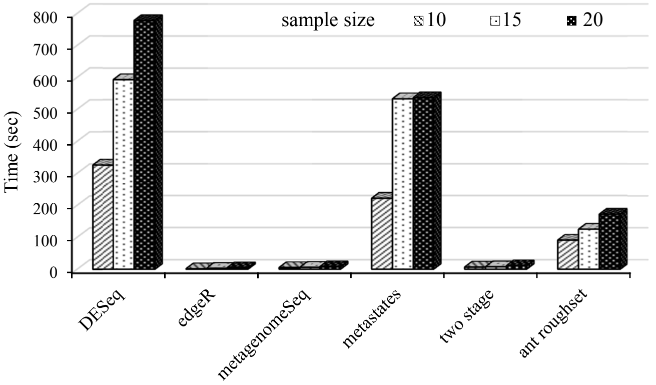


Fig. 2. Comparison of computational time (in second) for five methods under various sample sizes.

Table 2. Accuracy and Ψ_B for the presented method for sample size of 10, 15, and 20.

Sample size	Selected subset length	Accuracy (%)	Ψ	p-value
10	619	93	0.15	0.01074
15	527	97	0.17	0.001135
20	404	95	0.23	0.0002003

In Table 2, the accuracy Ψ_B and for the five methods are compared under various sample sizes. The subset obtained from the first phase in the proposed method is given as input to a SVM classifier and its accuracy is calculated. Regarding the Ψ_B values, we find that the precision criterion, accuracy, alone is not a suitable measure for evaluating the gene selection methods. So, in order to evaluate the methods, the Ψ_B criterion must be used. Also, the p-value shows that the accuracy obtained from the proposed method is not random, because its value is much lower than the usual 0.05 threshold; therefore, the result is reliable.

The ROC curve is usually used to measure signal detection. It is created by plotting the true positive rate (TPR) or sensitivity versus the false-positive rate (FPR) [25]. Figures 3 and 4 display the ROC curves of the proposed method before applying the gene selection process for sample sizes of 10 and 20, Figures 5 and 6 indicates the ROC curves after the gene selection process for samples 10 and 20. In general in ROC curves, the closer the curve is to the diameter the weaker the classifier is in distinction, and as the ROC curve tends to be upwards and far from the diameter, the classifier is better in distinction showing that method is better. According to Figures 5 and 6, the ROC curves plotted for the SVM classifier in sample sizes of 10 and 20 are far from the diameter, which shows a good performance of the classifier, and Figures 3 and 4 are closer to the diameter, which shows the classifier does not have proper efficiency. A better way to express this (near the curve to the diameter) is to have a surface below the ROC curve (AUC). On the other hand, as the AUC is closer to 0.5, the weaker

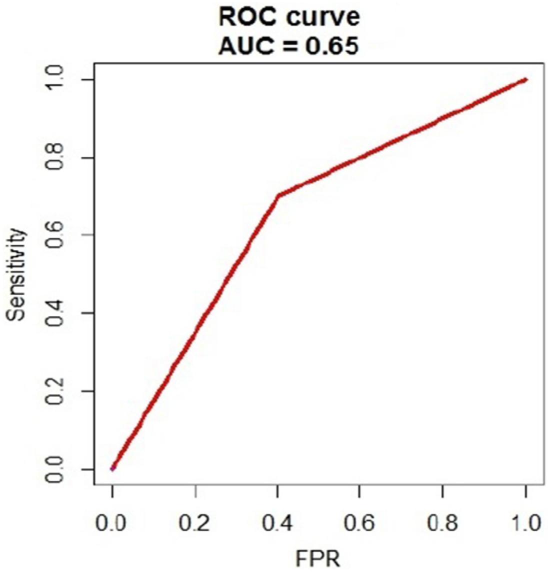


Fig. 3. ROC curves of the proposed method before applying the gene selection process for sample size of 20.

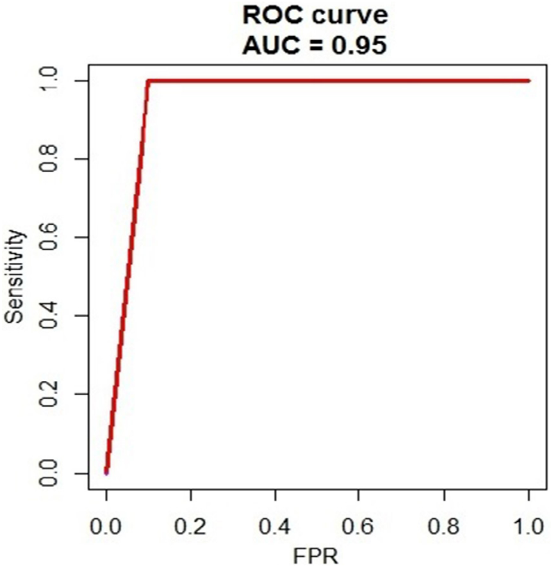


Fig. 5. ROC curves of the proposed method after applying the gene selection process for sample size of 20.

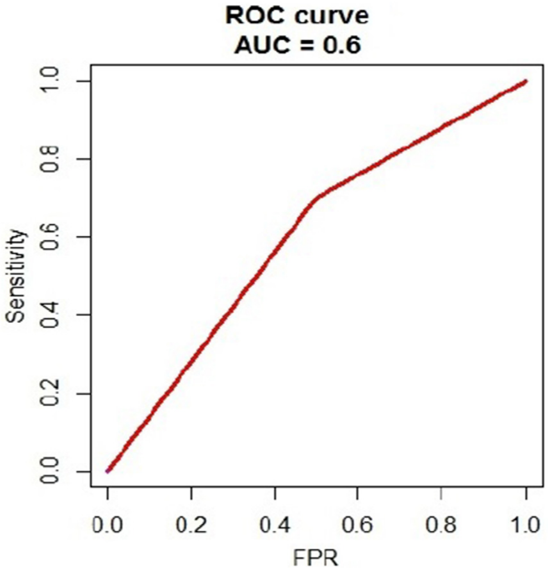


Fig. 4. ROC curves of the proposed method before applying the gene selection process for sample size of 10.

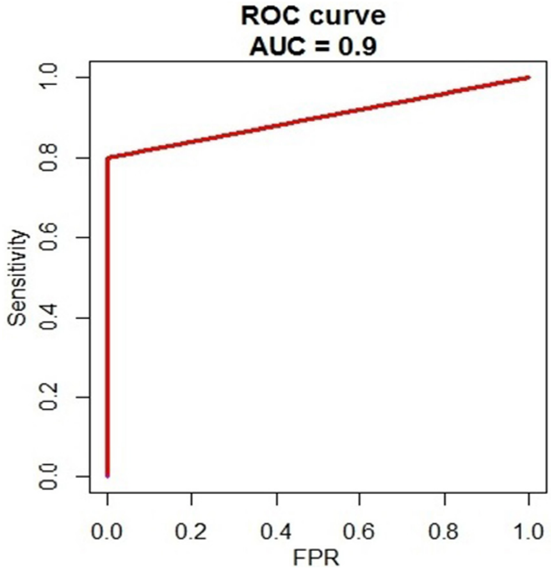


Fig. 6. ROC curves of the proposed method after applying the gene selection process for sample size of 10.

the classifier works in discrimination between the two groups, and whichever area is closer to one has a more favorable classification result. The AUC is independent of the various forms of population under investigation, and this is an advantage. In general, AUC represents the overall performance of the methods so that the greater the AUC value, the higher the performance of the method. Comparing the amount of AUC before and after the gene selection process, we found the overall performance of the both sample sizes 20 and 10, after the gene selection process, increased by 30%. It can also be realized that the number of samples has a significant effect on the classifier's performance, because a comparison of the ROC curve of the 20-sample size and the 10-sample size, is higher and further from the diameter. Figure 7 shows the results of the AUC derived from the five methods in various sample size. Typically, the AUC increases as the sample size increases.

As seen in Figure 7, the proposed method has a higher AUC than the rest of the methods and this means that the proposed method has a higher overall performance. In a sample size of 10 the proposed method increased by 2% compared to the two stage method. In a sample size of 15, the proposed method is 1% better than the two-stage method, and in a sample size of 20 the proposed method, in overall performance or AUC, had a 3% growth.

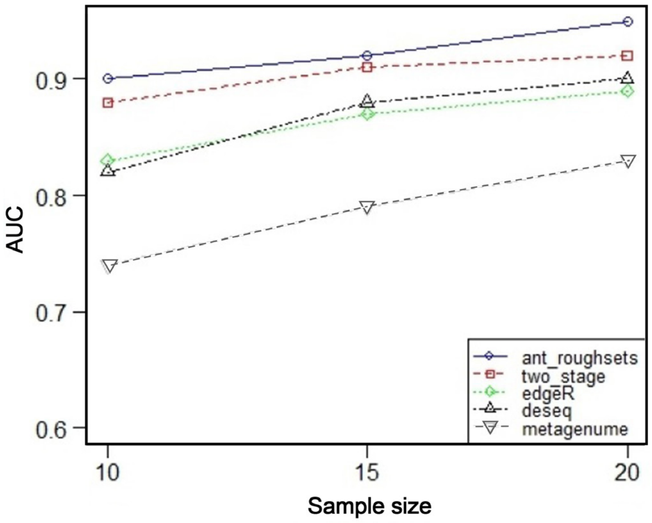


Fig. 7. Comparison of AUC curves.

Figures 8 and 9 indicate the precision and recall curves before applying the gene selection method, and Figures 10 and 11 show this curves after applying the gene selection method. In addition, this figure shows the plotted minimum and maximum PR curves and also a random PR curve to compare the performance of the method with the maximum, minimum and random mode. In this figure, the PR curve of the proposed method and the random PR curve are displayed by red and green, respectively. In general, precision and recall are opposite of each other but an ideal curve would have both equal to one. As seen in the curves, when the gene selection method is applied and the sample size increases the value of the recall increased, and when the precision rises the curve is more stable this means that the method has a higher performance. According to the curves, the amount of recall before applying the gene selection method in the sample size of 10 is approximately 0.59 and in the sample size of 20 it is nearly 0.63, which represents an increase of 4%. This amount of recall slowly diminished until the amount of precision reached a value of 1. The amount of recall, after applying a gene selection in a sample size of 10 is 1 and in the sample number of 20 it is approximately 90/0, this shows the recall is more stable in a sample size of 20. In addition, after selecting a gene, the PR curves do not have a significant difference between the maximum situations. Therefore, it can be concluded that classification is improved by applying the proposed method of gene selection.

Figure 12 shows the power in detection of the true differentially abundant genes for five methods at various levels of FDR for a sample size of 10. Typically, increasing the amount of FDR increases the detection power rate. According to the diagram, we find that the proposed method has a higher detection power in determining the genes which are statistically differentially abundant in two or more phenotypes. In general, in the proposed method the detection power in determining genes with differentially abundant for a sample size of 10, on average, has been increased by 1.3% compared to the two-stage method. Figure 13 displays the power in detection of the true differentially abundant genes for the five methods at various levels of FDR for a sample size of 20. FDR for a sample size of 20. In this diagram, as in the previous diagram, we find that the proposed method has the highest detection power. On average, in proposed method for sample size of 20 the detection power has been increased by 1% compared to the two-stage method.

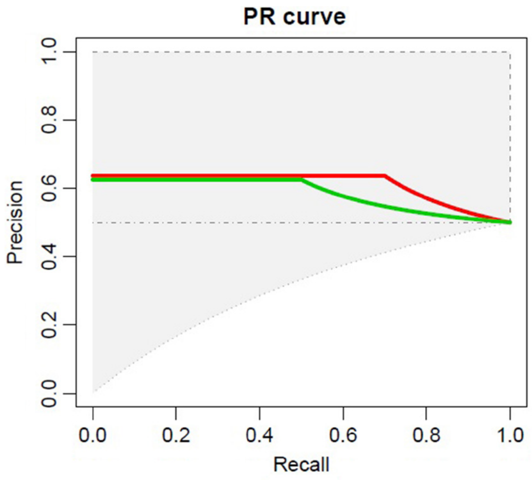


Fig. 8. The PR curves before applying the gene selection method for sample size of 10.

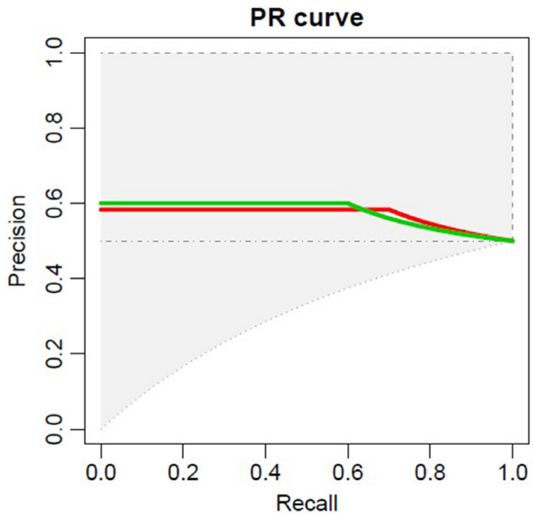


Fig. 9. The PR curves before applying the gene selection method for sample size of 20.

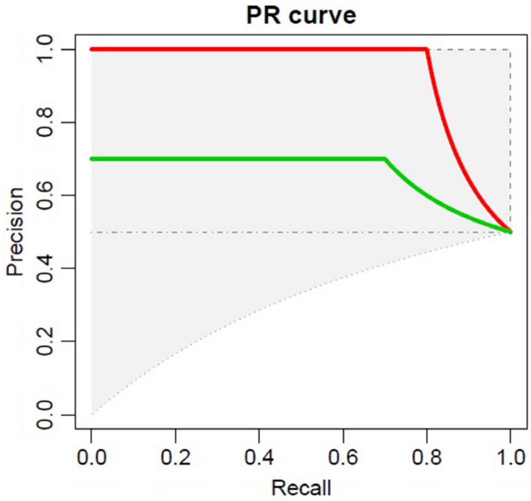


Fig. 10. The PR curves after applying the gene selection method for sample size of 10.

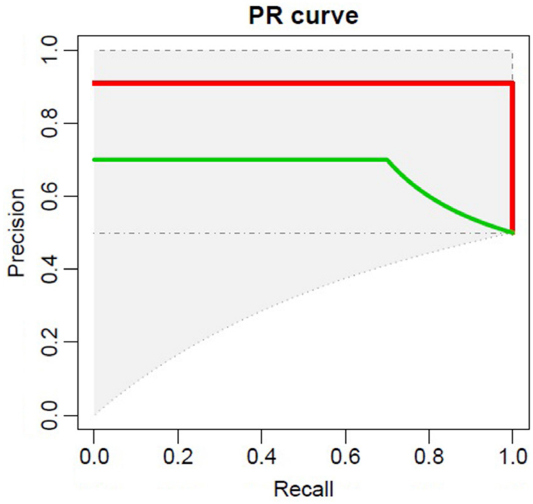


Fig. 11. The PR curves after applying the gene selection method for sample size of 20.

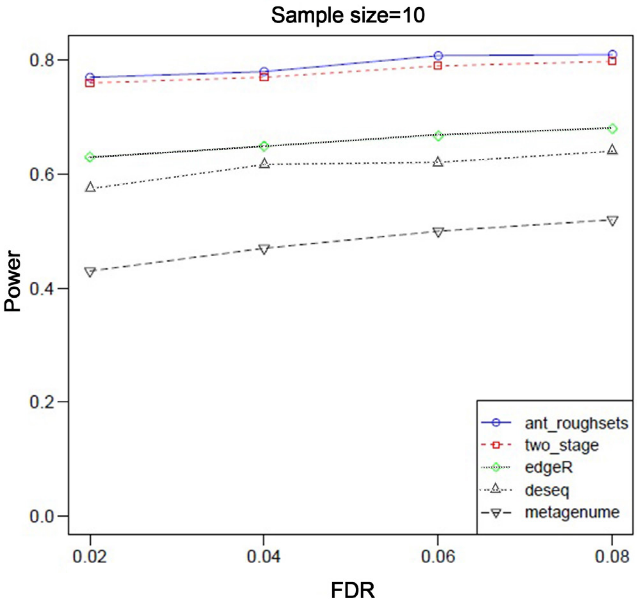


Fig. 12. The power in detection of the true differentially abundant genes for the five methods at various levels of FDR for sample size of 10.

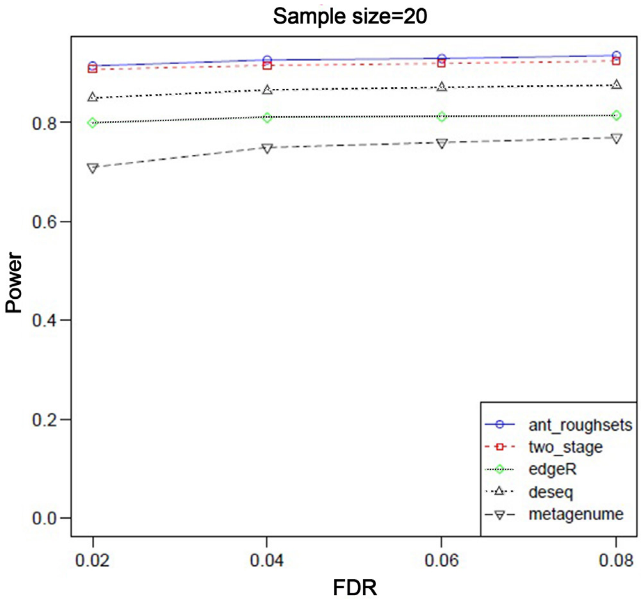


Fig. 13. The power in detection of the true differentially abundant genes for the five methods at various levels of FDR for sample size of 20.

6. Conclusion and future work

In this research, a gene selection method was proposed in the first stage to eliminate redundant and additional genes in a microarrays dataset. Then in the second stage, using the dataset obtained from the first step, the genes that differ in abundance in different phenotypes are identified to detect the presence of diseases. In the first step, a hybrid of the filter and wrapper gene selection method is introduced. In the filter section, genes are obtained using the ant colony algorithm and the information gain ratio which is calculated by fuzzy rough sets. Then the gene set obtained from the filter phase is evaluated in the wrapper section. Finally, the best subset of genes is collected. The gene set obtained from the first stage is used as the input of the second stage is used as the input of the second stage. Then the genes that are statistically differentially abundant in two or more phenotypes are identified using negative binomial distribution. The proposed method is implemented by the statistical software R. The results show that the proposed method is highly effective due to high accuracy, Ψ, ROC curves, and the increase of the AUC as compared to other existing methods, but this method has low runtime.

References

[1] V. Bolón-Canedo, N. Sánchez-Maroño, A. Alonso-Betanzos, J.M. Benítez, F. Herrera, A review of microarray datasets and applied feature selection methods, Inform. Sciences, 282 (2014) 111-135.

[2] H. Salem, G. Attiya, N. El-Fishawy, Classification of human cancer diseases by gene expression profiles, Appl. Soft Comput. 50 (2017) 124-134.

[3] P. Agarwalla, S. Mukhopadhyay, Bi-stage hierarchical selection of pathway genes for cancer progression using a swarm based computational approach, Appl. Soft Comput. 62 (2018) 230-250.

[4] H.H. Inbarani, A.T. Azar, G. Gothi, Supervised hybrid feature selection based on PSO and rough sets for medical diagnosis, Comput. Met. Prog. Bio. 113 (2014) 175-185.

[5] Y.Chen, Q.Zhu, H. Xu, Finding rough set reducts with fish swarm algorithm, Knowl-Based Syst. 81 (2015) 22-29.

[6] I.K. Park, G.S. Choi, Rough set approach for clustering categorical data using information-theoretic dependency measure, Inform. Syst. 48 (2015) 289-295.

[7] Z. Pawlak, A. Skowron, Rudiments of rough sets, Inform. Sciences, 177 (2017) 3-27.

[8] L.I. Kuncheva, Fuzzy rough sets: Application to feature selection, Fuzzy Set. Syst. 51 (1992) 147-153.

[9] R. Jensen, Q. Shen, Fuzzy-rough attributes reduction with application to web categorization, Fuzzy Set. Syst. 141 (2004) 469-485.

[10] M. Pradipta, G. Partha, Fuzzy-rough simultaneous attribute selection and feature extraction algorithm, IEEE T. Cybernetics, 43 (2013) 1166-1177.

[11] S. Zhao, E.C.C. Tsang, D. Chen, X. Wang, Building a rule-based classifier-a fuzzy rough set approach, IEEE T. Knowl. Data En. 22 (2010) 624-638.

[12] M. Dorigo, LM. Gambardella, A cooperative learning approach to the traveling salesman problem, IEEE T. Evolut. Comput. 1 (1997) 53-66.

[13] P. Schloss, J. Handelsman, Introducing SONS, a tool for operational taxonomic unit based comparisons of microbial community memberships and structures, Appl. Environ. Microb. 72 (2006) 6773-6779.

[14] B. Rodriguez-Brito, F. Rohwer, R.A. Edwards, An application of statistics to comparative metagenomics, BMC Bioinformatics, 7 (2006) 162.

[15] J. White, N. Nagarajan, M. Pop, Statistical methods for detecting differentially abundant features in clinical metagenomics samples, PLOS Comput. Biol, 5 (2009) e1000352.

[16] D. Huson, D. Richter, S. Mitra, A. Auch, S. Schuster, Methods for comparative metagenomics, BMC Bioinformatics, 10(Suppl 1) (2009) S12.

[17] Kristiansson, E. et al, ShotgunFunctionalizeR: An R-package for functional comparison of metagenomes, Bioinformatics, 25 (2009) 2737-2737.

[18] G.A. Montazer, S. ArabYarmohammadi, Detection of phishing attacks in Iranian e-banking using a fuzzy-rough hybrid system, Appl. Soft Comput. 35 (2015) 482-492.

[19] M. Podsiadło, H. Rybiński, Rough sets in economy and finance, In: Peters J.F., Skowron A. (eds) Transactions on Rough Sets XVII. Lecture Notes in Computer Science, Vol. 8375, pp. 109-173, 2014.

[20] C.H. Xie, Y.J. Liu, J.Y. Chang, Medical image segmentation using rough set and local polynomial regression, Multimed. Tools Appl. 74 (2015) 1885-1914.

[21] V. Prasad, T.S. Rao, M.S. Babu, Thyroid disease diagnosis via hybrid architecture composing rough data sets theory and machine learning algorithms, Soft Comput. 20 (2016) 1179-1189.

[22] M.P. Francisco, J.V. Berna-Martinez, A.F. Oliva, M.A.A. Ortega, Algorithm for the detection of outliers based on the theory of rough sets, Decis. Support Syst. 75 (2015) 63-75.

[23] J. Dai, Q. Xu, Attribute selection based on information gain ratio in fuzzy rough set theory with application to tumor classification, Appl. Soft Comput. 13 (2013) 211-221.

[24] M. Dorigo, L.M. Gambardella, A cooperative learning approach to the traveling salesman problem, IEEE T. Evolut. Comput. 1 (1997) 53-66.

[25] P. Naruekamol, M. Sohn, Q. Li, A two-stage statistical procedure for feature selection and comparison in functional analysis of metagenomes, Bioinformatics, 31 (2014) 157-165.