International Journal of Computational Intelligence and Informatics, Vol. 2: No. 2, July - September 2012

Performance Evaluation of Hybridized Rough Set based Unsupervised Approaches for Gene Selection

P. K. Nizar Banu
Department of Computer Applications
B. S. Abdur Rahman University
Chennai, India
nizarbanu@gmail.com

H. Hannah Inbarani
Department of Computer Science
Periyar University
Salem, India
hhinba@gmail.com

Abstract-Gene Selection aims to find a subset of highly informative genes from a problem domain which retains high accuracy to represent original genes. Rough Set Theory is adopted in this paper to discover the data dependencies and to reduce the number of genes contained in the dataset using the data alone without requiring additional information about the genes. Selecting genes in unsupervised learning scenarios is a harder problem than supervised gene selection due to the absence of class labels that would guide the search for relevant genes. PSO (Particle Swarm Optimization) is an evolutionary computation technique, which finds global optimum solution in many applications. This paper studies the performance of Unsupervised PSO based Relative Reduct (US-PSO-RR) and Unsupervised PSO based Quick Reduct (US-PSO-QR) approaches by applying it for a set of gene expression datasets to find the harmful genes easily. These two algorithms employs a population of particles existing within a multi-dimensional space and dependency measure that combines the benefits of both PSO and rough sets for better data reduction. The effectiveness of the algorithms is measured by using various clustering accuracy indices.

Keywords- Particle Swarm Optimization (PSO), quick reduct, relative reduct, rough sets, unsupervised gene selection.

I. INTRODUCTION

Selecting a subset of genes out of thousands of genes in a micro array data set without any information loss is very important to classify highly expressed genes and highly suppressed genes. The existing methods tries to select either all attributes or one attribute at a time which takes much time and it is computationally costly to select a subset of genes when it increases in dimension. This paper discusses the performance of rough set based unsupervised feature selection algorithms which adopts Particle Swarm Optimization technique that spawns in all possible directions at the very first time itself instead of scanning the genes in the database again and again as it is being done in the conventional feature selection methods.

Discriminant analysis is now being used in bioinformatics, such as distinguishing cancer tissues from normal tissues [2] or one cancer subtype vs. another [1]. Feature Selection (FS) [5] is defined as the problem of selecting more informative subset from a set of features based on some criterion. The genes removed should be noisy, redundant or of the least possible use. Genes selected should preserve the original meaning of the genes after reduction. A procedure which reduces the dimensionality using the existing information in the dataset and preserves the meaning of the genes is clearly desirable. Rough set can be used as a tool to discover data dependencies and to reduce the number of genes contained in a dataset using the data alone. Feature selection algorithms are classified as Filter approach and Wrapper approach. Filter based methods are efficient than wrapper based since it does not depend on any induction algorithm. Although wrappers produces good results, they are expensive to run, and will reduct more number of features. PSO, guides search to the optimal minimal subset every time is a heuristic filter-based method and is attractive for gene selection.

By using these approaches for selecting genes, the benefits of standard PSO and Rough sets is combined, the strict requirement of fitness function is relaxed, dependency among the attributes are introduced and are forced to select; thus a more flexible approach to predictive subset selection can be developed.

The rest of the paper is organized as follows. Section 2 focuses on research background. Section 3 discusses rough sets. Section 4 presents Unsupervised PSO based feature selection algorithms. Worked examples are presented in Section 5. Experimental results are presented and compared in section 6. Section 7 concludes the paper.

ISSN: 2349 - 6363

II. RESEARCH BACKGROUND

Two different types of approach to unsupervised feature selection have been adopted: those which maximize clustering performance using an index function [3], [11], and those which consider features for selection on the basis of dependency or relevance. The Quick Reduct (QR) algorithm given in [8], [14], attempts to calculate a reduct without exhaustively generating all possible subsets. The Relative Reduct (RR) algorithm, applied for gene selection is based on the measure of backward elimination of genes where attributes are removed from the set of considered genes if the relative dependency equals one upon their removal. Genes are considered one at a time, starting with the first, evaluating their relative dependency. One of the existing unsupervised feature selection algorithms USRR [16] calculates the dependency measures for every attribute. A new USQR algorithm is proposed in [17], introduces a new positive region based unsupervised subset evaluation measure using RST. In this method evaluation of the degree of dependency value for a feature's subset leads to each conditional attribute and evaluates mean of dependency values for all conditional attributes. In USQR and USRR methods, no decision attribute is required. If the relative dependency is equal to 1, even after removing a feature; then that feature can be removed. But it takes much iteration to converge which increases time and degrades the performance. With the help of these hybridized methods, high informative genes can be reducted in minimal iterations, and so it takes less time. Genes that are identical in many aspects are also considered as irrelevant and removed which results in information loss. Tolerance QuickReduct algorithm for supervised learning is proposed in [15]. This approach uses a threshold instead of checking the similarity to 1 exactly. The choice of threshold in this method permits attribute values to differ to a limited extent, allowing close values to be considered as identical. Unsupervised PSO based Quick Reduct (US-PSO-QR) is presented in [7] which starts by considering an empty set and then adds one feature at a time and checks the dependency between the features. It takes more number of iterations to converge and time as well. Unsupervised PSO based Relative Reduct (US-PSO-RR) is presented in [6] which starts by taking all the attributes as one set at a time and dependency measure is checked. If the dependency is not equal to 1, that attribute can be removed.

For dataset without decision class, both clustering or classification techniques are applied and the generated results such as clusters or classes are considered as decisions followed by feature selection methods. Unsupervised Feature Selection methods do not require any clustering or classification prior to gene selection.

This paper presents a study of unsupervised feature selection algorithms which integrates the merits of PSO and Rough sets.

III. ROUGH SETS

The rough set methodology was introduced by Pawlak [12] in the early 1980s as a mathematical tool to deal with uncertainty. It helps to discover data dependencies and to reduce the number of attributes, using the data alone. It requires no additional information [5], [10]. The basic concepts of the rough set theory and its philosophy are presented and illustrated with examples in the tutorial [18]. With the help of rough sets, irrelevant attributes can be removed with minimal information loss.

A. Lower and Upper Approximations

Let $I = (U, A)$ be an information system. Where *U* is the universe with a non-empty set of finite objects, *A* is a non-empty finite set of condition attributes. $\forall a \in A$, There is a corresponding function $f_a: U \rightarrow V_a$, where V_a, is the set of values of a. Let $X \subseteq U$, the P-lower approximation $\underline{P}X$ and P-upper approximation $\overline{P}X$ of a set *X* can be defined as:

$$\underline{P}\,X = \{\, x \in U | [x]_p \subseteq X \,\} \tag{1}$$

$$\overline{P}\,X = \{\, x \in U | [x]_p \cap X \neq \emptyset \,\} \tag{2}$$

Let $P, Q \subseteq A$ be an equivalence relations over U, and then the positive region can be defined as:

$$POS_P(Q) = \bigcup_{X \in U \,|\, Q} \underline{PX} \tag{3}$$

B. Relative Dependency Measure and Fitness Value

The unsupervised Relative Dependency measure for a particular particle is defined as follows:

$$\gamma_R(a) = \frac{|U \,/\, IND(R)|}{|U/IND(R \cup \{a\})|} \quad \forall a \notin R \tag{4}$$

Where *R* is the subset selected by the particle and the mean dependency of selected gene subset, on all the genes that are not selected by the particle is used as the fitness value of the particle X_i.

$$Fitness(X_i) = \overline{\gamma_R}(y) \; \forall y \notin R \tag{5}$$

IV. UNSUPERVISED PSO BASED FEATURE SELECTION ALGORITHMS

Particle Swarm Optimization (PSO) is an evolutionary computation technique [9]. The original idea was to graphically simulate the movement of bird flocking behavior. The authors of [13], introduced the concept of inertia weight into the particle swarm optimizer to produce the standard PSO algorithm.

A. Encoding

For applying US-PSO-QR and US-PSO-RR algorithms, each particle's position is represented as binary bit strings of length N, where N is the total number of particles [13]. Each particle's position is an attribute subset.

For example, if *a, b, c* and d are attributes and if the selected random particle is *(1, 0, 0, 1)*, then the attribute subset is *(a, d)*.

a	**b**	**c**	**d**
1	0	0	1

B. Representation and Updation of Velocity and Positions

Each particle's velocity is represented as a positive integer varying between 1 and V_{max}. It implies how many of the particle's bit should be changed as that of the global best position.

For selecting genes, PSO is initialized with a population of particles. Each particle is treated as a point in an S-dimensional space. The i^{th} particle is represented as $X_i = (x_1, x_2, \dots, x_{is})$. The best position of any particle *pbest* of any particle $P_i = (p_{i1}, p_{i2}, \dots, p_{is})$. The index of the global best particle is represented by *gbest*. The velocity of a particle is $V_i = (v_{i1}, v_{i2}, \dots v_{is})$. The particles velocity and position are updated as follows:

Each particle's velocity is updated using:

$$V_{id} = w * v_{id} + c_1 * rand() * (p_{id} - x_{id}) + c_2 * Rand() * (p_{gd} - x_{id}) \tag{6}$$

where *w* is the inertia weight, c_1 and c_2 are acceleration constants. Based on velocity, Particle's position is updated as follows:

- If $V \leq xg$, randomly change *V* bits of the particle, which are different from that of *gbest.*
- If $V > xg$, change all the different bits to be the same as that of *gbest* and a further (V-xg) bits should be changed randomly.

w can be calculated as follows

$$w = w_{max} - \frac{w_{max} - w_{min}}{iter_{max}} \; iter \tag{7}$$

Here w_{max} is the initial value of the weighting coefficient, w_{min} is the final value, $iter_{max}$ is the maximum number of iterations and iter is the current iteration.

Procedure for selecting genes using the US-PSO-RR algorithm given in figure 1 is described as follows. This algorithm calculates a reduct set without generating all possible subset. It starts by selecting random values for each particle and velocity. A population of particles is constructed with random positions and velocities on S dimensions in the problem space. For each particle X_i, 1's are taken as the selected genes and 0's are considered as removed genes. The average dependency of selected genes on each non-selected gene is computed. If the mean dependency is equal to 1, then the gene subset of the particle is considered as the reduct set. If the mean dependency is not equal to 1, the *Pbest* (highest relative dependency value) of each particle is retained and the best value of the entire population is retained as the global best value. Then the position and velocity are updated as defined above and the next population is generated and the fitness values are computed for each particle until fitness value of the selected gene subset becomes 1.

Procedure for selecting genes using the US-PSO-QR algorithm given in Figure 2 is described as follows. This algorithm calculates a reduct set without generating all possible subset. It starts with an empty set and it adds one attribute at a time, in turn. A population of particles is constructed with random positions and velocities on S dimensions in the problem space. Fitness function for each particle which is represented as 1, is evaluated using the following equation.

Algorithm : US-PSO-RR (C)
Input : C, the Set of all conditional features
Output : Reduct R
Step 1: Initialize X with random position and V_i with random velocity
∀: $X_i \leftarrow$ randomPosition (); $V_i \leftarrow$ randomVelocity(); fit ← 0; globalbest ← fit;
gbest ← X_1 ; pbest(1) ← X_1
For i = 1 . . . S
pbest(i) = X_i; Fitness(i) =0
End For
Step 2: While Fitness != 1 // Stopping Criterion
For i = 1, . . . ,S // for each particle
∀: X_i; Compute fitness of feature subset of X_i
R ← Feature subset of X_i (1's of X_i)
$\forall a \in y$

$$\gamma_R(a) = \frac{|U / IND(R)|}{|U / IND(R \cup \{a\})|}$$

Fit = $\overline{\gamma_R}(y)\ \forall y \notin R$
If Fitness(i) >fit
Fitness(i) = fit
pbest(i) = X_i
End
If (Fit == 1)
return R
End if
End For
Step 3: Compute best fitness
For i = 1, . . . , S
If (Fitness (i) > globalbest)
gbest ← X_i; globalbest ← Fitness(i); gbest ← X_i;
End if
End For
UpdateVelocity (); // Update Velocity V_i's of X_i's
UpdatePosition (); // Update position of X_i's, Continue with the next iteration
End {while}
Output Reduct R

Figure 1. US-PSO-RR Algorithm

$$Fitness(i) = \overline{\gamma_{T \cup \{x\}}}(y)\ \forall y \in C \quad (8)$$

Where $\gamma_{T \cup \{x\}}(y) = \frac{|POS_{T \cup \{x\}}(y)|}{|U|}$ (9)

A gene with highest fitness value is taken and all possible combinations of the selected gene with the other genes are constructed. Fitness of the selected genes with different combinations is calculated. If the current particle's fitness evaluation is better than the *Pbest*, then this particle becomes the current best and its position and fitness are stored. Then, the current particle's fitness is compared with population's overall previous best fitness. If the current value is better than *gbest*, then this is set to the current particle's position, with the global best fitness updated. This position represents the best feature subset encountered so far, and is stored in *R*. The velocity and position of the particle is then updated. This process is carried out until the stopping criterion is met, usually a maximum number of iterations. The returned feature subset is a US-PSO-QR set. According to the algorithm, the mean dependency of each attribute subset is calculated and the best particle is chosen.

V. WORKED EXAMPLE

A. Unsupervised PSO based Relative Reduct (US-PSO-RR)

To illustrate the operation of the US-PSO-QR and US-PSO-RR algorithms, it is applied to the example data set given in Table 1, which contains four conditional attributes. The population generated initially is given in Table 2.

Algorithm : US-PSO-QR (C)
Input : C, the set of features
Output : Reductset R

Step 1: Initialize X with random position and V_i with random velocity
$\forall: X_i \leftarrow$ randomPosition (); $V_i \leftarrow$ randomVelocity(); fit ← 0; globalbest ← fit;
gbest ← X_1
Step 2: While Fitness $!= \overline{\gamma_C}(y)\ \forall y \in C$ // Stopping criterion
For i = 1, . . . , S // for each particle
$\forall: X_i$; T ← { }
Compute fitness of feature subset of X_i
R ← Feature subset of X_i (1's of X_i)
$\forall x \in R;\ \forall y \in C$

$$\gamma_{T \cup \{x\}}(y) = \frac{\left|POS_{T \cup \{x\}}(y)\right|}{|U|}$$

$$\text{Fitness(i)} = \overline{\gamma_{T \cup \{x\}}}(y)\ \forall y \in C$$

End For
Step 3: Compute best fitness
For i = 1, . . . , S
If (fitness(i) > globalbest)
$g$best ← X_i; globalbest ← Fitness(i); pbest(i) ← bestPos(X_i); gbest ← X_i;
If fitness(i) = $\overline{\gamma_C}(y)\ \forall y \in C$
R ← getReduct(X_i)
End if
End if
End For
UpdateVelocity(); // Update Velocity V_i's of X_i's
UpdatePosition(); // Update Position of X_i's, Continue with the next iteration
End {while}
Output Reduct R

Figure 2. US-PSO-QR Algorithm

Table 1. Example Dataset

$x \in U$	a	b	c	d
1	1	0	2	1
2	1	0	2	0
3	1	2	0	0
4	1	2	2	1
5	2	1	0	0
6	2	1	1	0
7	2	1	2	1

Table 2. Population Generated

Particles	Position			
X1	1	0	0	1
X2	0	1	0	1
X3	1	1	0	0
X4	0	1	1	0

For Particle X_1,

Initial population generated is *(1, 0, 0, 1)*, Feature subset selected is *(a, d)*, Feature subset removed is *(b, c)*. Hence $R = \{a, d\}, Y = \{b, c\};\ \forall a \in Y$

$$\gamma_R(b) = \frac{|IND_R|}{|IND_{R\cup\{b\}}|} = \frac{\{1,4\}\{2,3\}\{5,6\}\{7\}}{\{1\}\{2\}\{3\}\{4\}\{5,6\}\{7\}} = \frac{4}{6} = 0.667$$

$$\gamma_R(c) = \frac{|IND_R|}{|IND_{R\cup\{b\}}|} = \frac{\{1,4\}\{2,3\}\{5,6\}\{7\}}{\{1,4\}\{2\}\{3\}\{5\}\{6\}\{7\}} = \frac{4}{6} = 0.667$$

$$\overline{\gamma_R}(a)\forall a \in Y = \frac{(0.667+0.667)}{2} = 0.667$$

Since relative dependency $\overline{\gamma_R}(a) \neq 1$, $\{a, d\}$ is not the selected feature subset. Proceeding in this way, after few iterations, Population like $X_i = \{0, 1, 1, 1\}$ is generated and the dependency is calculated as follows

$R= \{b, c, d\}$, $Y= \{a\}$

$$\gamma_R(a) = \frac{|IND_R|}{|IND_{R\cup\{a\}}|} = \frac{\{1\}\{2\}\{3\}\{4\}\{5\}\{6\}\{7\}}{\{1\}\{2\}\{3\}\{4\}\{5\}\{6\}\{7\}} = \frac{7}{7} = 1$$

$$\overline{\gamma_R}(a) \quad \forall a \in Y = 1$$

Since, there is only one attribute in *Y*.

B. Unsupervised PSO based Quick Reduct (US-PSO-QR)

For Particle X_1, initially it starts with $T \leftarrow \{\ \}$, then proceeds as follows

$$\gamma_{T\cup\{ad\}}(a) = \frac{|POS_{T\cup\{ad\}}(a)|}{|U|} = \frac{|\{1,2,3,4,5,6,7\}|}{|\{1,2,3,4,5,6,7\}|} = \frac{7}{7} = 1$$

$$\gamma_{T\cup\{ad\}}(b) = \frac{|POS_{T\cup\{ad\}}(b)|}{|U|} = \frac{|\{5,6,7\}|}{|\{1,2,3,4,5,6,7\}|} = \frac{3}{7} = 0.4286$$

$$\gamma_{T\cup\{ad\}}(c) = \frac{|POS_{T\cup\{ad\}}c|}{|U|} = \frac{|\{1,4,7\}|}{|\{1,2,3,4,5,6,7\}|} = \frac{3}{7} = 0.4286$$

$$\gamma_{T\cup\{ad\}}(d) = \frac{|POS_{T\cup\{ad\}}(d)|}{|U|} = \frac{|\{1,2,3,4,5,6,7\}|}{|\{1,2,3,4,5,6,7\}|} = \frac{7}{7} = 1$$

$$\overline{\gamma_{T\cup\{ad\}}}(y);\ \forall y \in C = \frac{1 + 0.4286 + 0.4286 + 1}{4} = \frac{2.8572}{4} = 0.7143$$

$$\overline{\gamma_{T\cup\{bd\}}}(y);\ \forall y \in C = 0.9286$$

$$\overline{\gamma_{T\cup\{ab\}}}(y);\ \forall y \in C = 0.5714$$

$$\overline{\boldsymbol{\gamma}_{\mathbf{T}\cup\{\mathbf{bc}\}}}(\mathbf{y});\ \forall \mathbf{y} \in \mathbf{C} = \mathbf{0.9286}$$

It proceeds like this, using global optimization method. It explores in all possible directions. It randomly selects the local and global optimization and converges easily. Since, the dependency of attributes is not equal to 1, next iteration is carried out. In the second iteration, randomly selected particles are *b, c and d.* It is denoted as *(0, 1, 1, 1).*

$$\gamma_{T\cup\{bcd\}}a = \frac{|POS_{T\cup\{bcd\}}a|}{|U|} = \frac{|\{1,2,3,4,5,6,7\}|}{|\{1,2,3,4,5,6,7\}|} = \frac{7}{7} = 1$$

$$\gamma_{T\cup\{bcd\}}b = \frac{|POS_{T\cup\{bcd\}}b|}{|U|} = \frac{|\{1,2,3,4,5,6,7\}|}{|\{1,2,3,4,5,6,7\}|} = \frac{7}{7} = 1$$

$$\gamma_{T\cup\{bcd\}}c = \frac{|POS_{T\cup\{bcd\}}c|}{|U|} = \frac{|\{1,2,3,4,5,6,7\}|}{|\{1,2,3,4,5,6,7\}|} = \frac{7}{7} = 1$$

$$\gamma_{T\cup\{bcd\}}d = \frac{|POS_{T\cup\{bcd\}}d|}{|U|} = \frac{|\{1,2,3,4,5,6,7\}|}{|\{1,2,3,4,5,6,7\}|} = \frac{7}{7} = 1$$

$$\overline{\gamma_{T \cup \{bcd\}}}(y);\ \forall y \in C = 1$$

The subset "*b, c, d*" produces the dependency 1.

VI. EXPERIMENTAL RESULTS

In this section, the US-PSO-RR algorithm and US-PSO-QR algorithms are studied, compared and evaluated using clustering indices.

A. Data Sets

Two datasets namely Leukemia and Lung Cancer gene expression datasets which are available in the website http://datam.i2r.a-star.edu.sg/datasets/krbd/ [19] are taken for experimentation. The details of the dataset used are given in Table 3.

Table 3. Details of Gene Expression Datasets

Dataset	Number of Genes	Class	Number of Samples
Leukemia	7129	ALL / AML	34 (20 / 14)
Lung Cancer	7129	Tumor / Normal	96 (86 / 10)

The data presented in Table 4, presents the number of genes selected by the US-PSO-RR and US-PSO-QR algorithms.

Table 4. Genes selected by US-PSO-RR and US-PSO-QR

Index	Data set	Objects	Attribute Size	US-PSO-RR	US-PSO-QR	No. of Att. reducted in common
				No. of Att. Reducted	No. of Att. Reducted	
1	Leukemia	34	7129	3606	3558	2453
2	Lung Cancer	96	7129	3621	3607	2762

B. Comparison of US-PSO-RR and US-PSO-QR Methods

Genes are selected by the feature selection algorithms and are clustered in to two clusters. The genes that causes tumor are grouped in one cluster whereas the harmless genes are grouped in the other cluster. Dataset with class attribute is taken for our experiment to check the effectiveness of the feature selection algorithms. The feature selection algorithms discussed in this paper are applied without considering the class attribute. Clustering was performed on the reduced datasets which were obtained using the Unsupervised PSO based Relative Reduct (US-PSO-RR) and Unsupervised PSO based Quick Reduct (US-PSO-QR) methods. Results are presented in terms of clustering accuracy. To evaluate the performance of the feature selection algorithms we used Mean Absolute Error (MAE), Root Mean Square Error (RMSE) [4] and Xie-Beni validity which measures the compactness and separation of clusters [20]. Table 5 and 6 shows the Rough K Means clustering performance of features selected using US-PSO-RR and US-PSO-QR.

Table 5. Rough K Means Clustering Performance Rate for the Features Selected using Unsupervised PSO based Relative Reduct (US-PSO-RR)

Data Set	Objects	Rough K Means for US-PSO-RR		
		RMSE	MAE	Xie-Beni
Leukemia	7129	0.0296	0.0871	0.1254
Lung Cancer	7129	0.0241	0.0641	0.3171

Table 6. Rough K Means Clustering Performance Rate for the Features Selected Using Unsupervised PSO based Quick Reduct (US-PSO-QR)

Data Set	Objects	Rough K Means for US-PSO-QR		
		RMSE	MAE	Xie-Beni
Leukemia	7129	0.0963	0.2172	0.6083
Lung Cancer	7129	0.0227	0.0677	0.2152

Table 7 and 8 shows the K Means clustering performance of the features selected using US-PSO-RR and US-PSO-QR.

Table 7. K Means Clustering Performance Rate for the Features Selected using Unsupervised PSO based Relative Reduct (US-PSO-RR)

Data Set	Objects	K Means for US-PSO-RR		
		RMSE	MAE	Xie-Beni
Leukemia	7129	0.0883	0.0246	0.4013
Lung Cancer	7129	0.0241	0.0641	0.2659

Table 8. K Means Clustering Performance Rate for the Features Selected using Unsupervised PSO based Quick Reduct (US-PSO-QR)

Data Set	Objects	K Means for US-PSO-QR		
		RMSE	MAE	Xie-Beni
Leukemia	7129	0.0364	0.1087	0.1076
Lung Cancer	7129	0.0507	0.1020	0.2342

Figure 3, 4 and 5 shows the comparative analysis of RMSE, MAE and Xie-Beni index of Rough K Means clustering for the genes selected by US-PSO-RR and US-PSO-QR. It is observed that Unsupervised PSO based Relative Reduct (US-PSO-RR) approach selects more suitable genes for further medical analysis than the Unsupervised PSO based Quick Reduct (US-PSO-QR) approach.

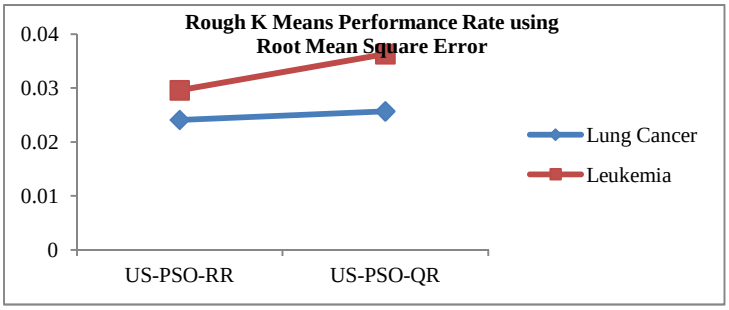

Figure 3. Rough K Means Performance Rate using Root Mean Square Error

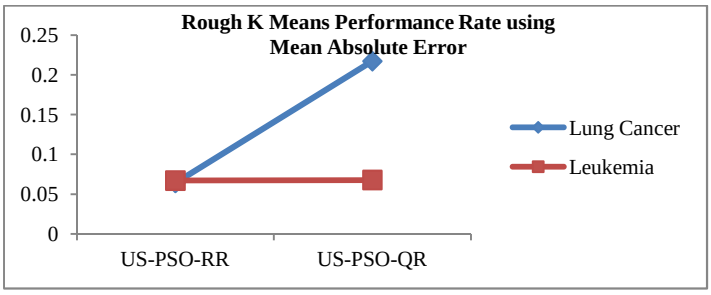

Figure 4. Rough K Means Performance Rate using Mean Absolute Error

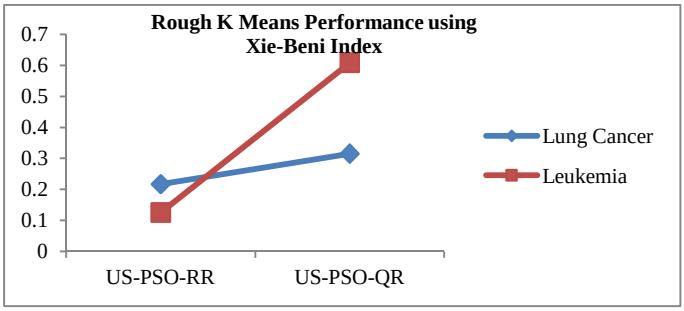

Figure 5. Rough K Means Performance Rate using Xie-Beni Index

Figure 6, 7 and 8 shows the comparative analysis of RMSE, MAE and Xie-Beni index of K Means clustering for the genes selected by US-PSO-RR and US-PSO-QR. It is observed that K Means also produces more suitable genes selected by Unsupervised PSO based Relative Reduct (US-PSO-RR) for medical analysis than the genes selected by Unsupervised PSO based Quick Reduct (US-PSO-QR) approach.

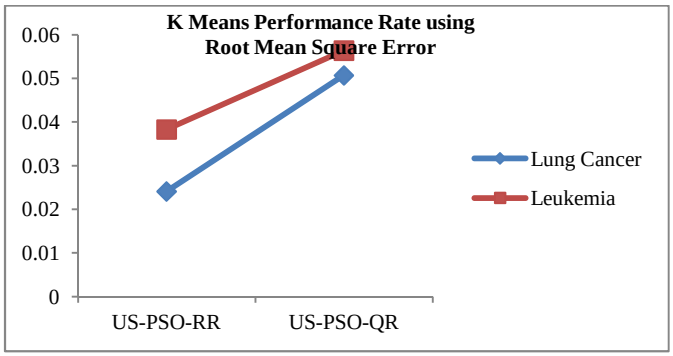

Figure 6. K Means Performance Rate using Root Mean Square Error

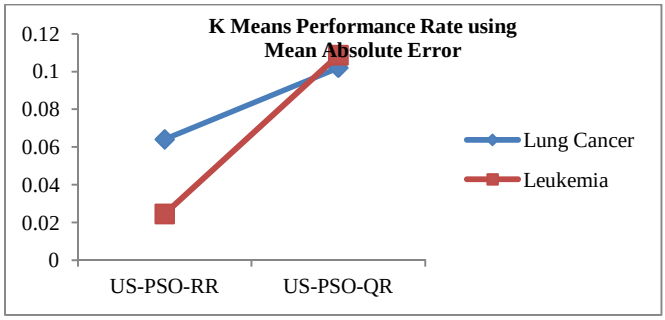

Figure 7. K Means Performance Rate using Mean Absolute Error

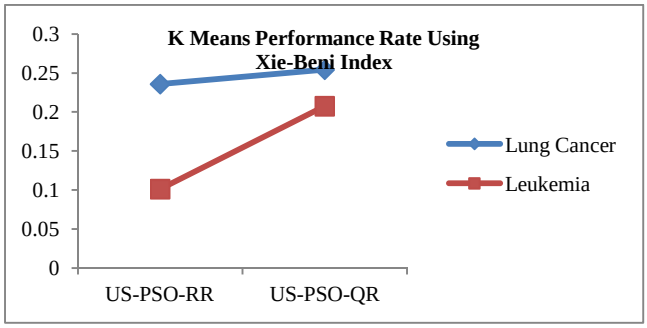

Figure 8. K Means Performance Rate using Xie-Beni Index

VII. Conclusion and Future Enhancement

This section presents the results of the US-PSO-RR and US-PSO-QR algorithms. Comparison is made to suggest the more suitable genes for medical diagnosis. The results are presented in terms of reduct set, Rough K Means and K Means clustering performances. Experimental results for two gene expression data sets are presented. Unlike the other existing unsupervised feature selection methods, which starts either with an empty set or with a set of all features, US-PSO-RR and US-PSO-QR approaches selects random particles and explores in all possible directions thus converges in global optimization. These approaches are highly suitable for large datasets. The efficiency of the two feature selection approaches discussed in this paper is clearly exhibited. Although the US-PSO-RR approach selects more genes than the US-PSO-QR approach the clustering performance result shows US-PSO-RR approach is better than US-PSO-QR approach. The empirical results also reveal the importance of using US-PSO-RR for datasets which has no decision attributes. In future the same approach can be extended to medical image datasets for cancer diagnosis.

REFERENCES

[1]. Alizadeh A.A., Eisen MB, Davies RE, Ma C, Lossos IS Rosenwald A, Boldrick JC, Saber H, Tran T, Yux et al. "Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling", Nature 403 (6769), 2000, pp.503-511, DOI:10.1038/35000501.

[2]. U. Alon, N. Barkai, D. A. Notterman, K. Gish, S. Ybarra, D. Mack, and A. J. Levine, "Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays", PNAS vol. 96, no. 12, pp. 6745-6750, 1999.

[3]. M. Dash, and H. Liu., "Unsupervised feature selection", Proc. Pacific and Asia Conf. Knowledge Discovery and Data Mining, Kyoto, 2000, pp. 110–121.

[4]. O. F. De Franca, H. M. Ferreira, and F. J. Von Zuben, "Applying Biclustering to Perform Collaborative Filtering", Proc. Seventh International Conference on Intelligent Systems Design and Applications. 2007, pp. 421-426.

[5]. J. W. Grzymala-Busse, "LERS-a system for learning from examples based on rough sets", Intelligent Decision Support, Springer, 1992. pp. 3–18.

[6]. H. Hannah Inbarani, and P. K. Nizar Banu, "Unsupervised Hybrid PSO – Relative Reduct Approach for Feature Reduction", Proc. International Conference on Pattern Recognition, Informatics and Medical Engineering (PRIME), March 2012, pp. 103-108, DOI: 10.1109/ICPRIME.2012.6208295.

[7]. H. Hannah Inbarani, P. K. Nizar Banu, S. Andrews, "Unsupervised hybrid PSO - Quick reduct approach for feature reduction", Proc. International Conference on Recent Trends In Information Technology (ICRTIT), April 2012, pp. 11-16. DOI: 10.1109/ICRTIT.2012.6206775.

[8]. R. Jensen, and Q. Shen, "Fuzzy-rough attribute reduction with application to web categorization", Fuzzy Sets and Systems, vol. 141, no. 3, 2004, pp. 469–485, DOI: 10.1016/S0165-0114(03)00021-6.

[9]. J. Kennedy, and R. Eberhart, "Swarm Intelligence", Morgan Kaufmann, 2001.

[10]. P. Mitra, C. A. Murthy, and S. K. Pal, "Unsupervised feature selection using feature similarity," IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 24, no. 4, 2000, pp. 1–13.

[11]. S. K. Pal, R. K. De, and J. Basak., "Unsupervised feature evaluation: a neuro-fuzzy approach". IEEE Trans. on Neural Networks, vol. 11, no. 2, 2000, pp. 366–376, DOI: 10.1109/72.839007.

[12]. Z. Pawlak, "Rough sets", International Journal of Computer and Information Science, vol. 11, pp. 341–356, 1982.

[13]. Richard Jenson, and Qiang Shen, "Rough Set-Based Feature Selection – A Review", in Rough Computing, Theories, Technologies and applications, IGI Global, USA, 2008, ch. 3, pp. 70 – 107.

[14]. Shi, Y., Eberhart, R.C., "Parameter selection in particle swarm optimization". Proc. Evolutionary Programming VII (EP98), New York: Springer-Verlag 1998, pp. 591–600.

[15]. A. Skowron, and J. Stepaniuk, "Tolerance Approximation Spaces", Fundamenta Informaticae, vol. 27, no. 2, 1996, pp. 245–253.

[16]. C. Velayutham, and K. Thangavel, "Unsupervised Feature Selection using Rough Set", Proc. International Conference on Emerging Trends in Computing, March 2011, pp. 307-314.

[17]. C. Velayutham, and K. Thangavel, "Unsupervised Quick Reduct Algorithm Using Rough Set Theory", Journal of Electronic Science and Technology, vol. 9, no. 3, pp. 193-201, September 2011.

[18]. B. Walczak, and D. L. Massart, "Rough sets theory", Chemometr. Intell. Lab. Syst. Vol. 47 No. 1, 1999, pp. 1 – 16.

[19]. Xiaosheng Wang, Osamu Gotoh, "Cancer Classification Using Single Genes", pp 179-188.

[20]. X. L. Xie., and G. Beni, "A validity measure for fuzzy clustering", IEEE Trans, Pattern Analysis and Machine Intelligence, vol. 13. No. 8, 1991, pp. 841-847, DOI: 10.1109/34.85677.