Comparison of Feature Selection Techniques for Thyroid Disease

Surekha S[1], JayaSuma G[2]

***Abstract*—** Past decades, Rough Set Theory (RST) has been one of the efficient and effective methods for Feature Subset Selection (FSS). Even though RST is a popular method for FSS, could not generate minimal subset of features with acceptable accuracy. Therefore, Evolutionary Computation (EC) techniques have been proposed to find the minimal subset of features with significant accuracy for a particular task. In this paper, the performance of two EC algorithms; Particle Swarm Optimization (PSO) and Genetic Algorithm (GA) are compared with Quickreduct and Johnson's reduct computation algorithms of RST. To evaluate the performance of the EC and RST approaches, reduced dataset has been applied to three classifiers namely Decision tree, Naïve Bayesian and K-Nearest Neighbor for classification of the Thyroid disease. The experimental results showed that EC approach for FSS outperforms RST based approach in both the terms of generating minimal subset of features and resulting in higher classification accuracy.

***Keywords*—**Classification, EvolutionaryComputationTechniques, Feature Subset Selection, Rough Set Theory.

I. Introduction

As information technology is growing rapidly, the complexity and dimensionality of real world datasets is affecting the accuracy of the classification. All the attributes in the dataset will not contribute for a particular task; an efficient and effective mechanism is required to separate relevant attributes from irrelevant attributes. So from the huge datasets most relevant attributes are to be extracted, that contributes the maximum to the decision and also retains the suitable accuracy of the classification techniques. The separation of the relevant from irrelevant attributes is known as FSS [1]. FSS plays an indispensable role in the research of Knowledge acquisition [2] due to the abundance of noisy and superfluous features. Too many irrelevant and superfluous features are increasing the demand for excessive memory requirements and reducing the prediction accuracy of the classification technique. In order to overcome the difficulties with massive datasets, determine a minimal subset of features while maintaining or to improve the accuracy as in representing the original set of features. The existing FSS algorithms are broadly classified into two categories namely Filter approaches and Wrapper approaches [3] based on whether they employed a learning algorithm during the feature selection process. The Wrapper approach finds and evaluates the feature subset based on a learning algorithm whereas the filter approach generates a minimal subset independent of the classifier. In recent years, RST has been used as a filter approach for the FSS. RST selects the predominant attributes of the given dataset without modifying the data during the process of selecting the relevant attributes. Most of the researchers stated that there are some disadvantages with RST based reduct computation algorithms like Johnson's and Quickreduct. EC algorithms [4] have been introduced and succeeded in finding relevant features with minimum subset while retaining the prediction accuracy of a particular data mining task. Now-a-days classification has became an important data mining task in the area of medical diagnosis, and the various classification methods are Decision Tree, Rule-based classifier, Naïve Bayesian, K-Nearest Neighbor, and Support Vector Machines. Decision Tree classifier is simple and is easy to understand, Naïve Bayesian classifier requires very less information (probabilities) about the data to complete the classification task, and the K-Nearest Neighbor is a lazy learner which takes very less time to learn from the training data. In this study, two popular EC algorithms GA and PSO are applied to Thyroid dataset for generating minimal subset of features with increased accuracy. GA[5] is a filter approach and PSO[6] is a wrapper approach for FSS. In medical perspective, the FSS helps to identify minimal set of features that are most effective and informative for the diagnosis of diseases with very few tests.

In this paper, Section.II summarizes the related work of FSS and classification of various medical datasets. Section.III gives an introduction to the RST by discussing various reduct computation algorithms focusing on dimensionality reduction and Section.IV gives a brief introduction of the EC approach for FSS and discusses the GA and PSO. Section.V presents the methodology. The analyses of the experimental results are discussed in Section.VI and Section VII gives the conclusion and future work

II. Related Work

Kohavi *et al.*[3] applied sequential forward selection and sequential backward selection wrapper feature selection techniques for finding best features, but they are affected by the local optima and they require high computational time for finding reducts. Starzyk *et al.*[7] generated reducts by the simplification of discernibility function. Yao *et al.*[8] applied rough set concepts i.e., discernibility matrix and discernibility

Surekha S[1] is Research Scholar, Computer Science and Engineering, Jawaharlal Nehru Technological University Kakinada, Kakinada, Andhra Pradesh, India. (e-mail: surekha.samsani@gmail.com).

JayaSuma G[2], is Head of the Department IT, JNTUK-UCEV, Vizianagaram, Andhra Pradesh, India (e-mail:gjscse@gmail.com).

function for attribute reduction. W'roblewski *et al.*[5] applied genetic algorithms to generate minimal reduct set. But it is based on distinction table which requires high processing time. Choudary *et al.* [9] studied and compared three reduct computation algorithms Exhaustive, Johnson, Quickreduct and Genetic algorithm for generating the rules for the diagnosis of Breast Adenocarcinoma using RST. Kennedy *et al.*[6] introduced particle swarm optimization for neural networks. Chuang *et al.* [10] used the catfish effect to Particle Swarm Optimization for feature selection, which introduces the new particles by replacing the worst particles when there is no improvement in gbest for a pre-determined number of iterations. Xue *et al.* [11] developed new initialization and updating mechanisms for pbest and gbest in PSO for feature selection, which increased the classification accuracy by reducing the number of features and computation time.

III. Rough Set Theory

RST, a mathematical tool based on approximations was firstly proposed by the Polish scientist Z.Pawlak during the early 1980's. RST [12-15] provides efficient methods to discover hidden patterns in data by identifying the dependencies in data. RST helps us to find out the minimal attribute sets with minimum information loss and using the data alone i.e., requiring no additional information about data like probability or membership functions in the Fuzzy Set theory [16]. The basic concepts of RST are given below.

A. Information System

Let IS=(U,A) represents an Information System [17], where U is the nonempty finite set of objects called the Universe and A is a nonempty finite set of attributes such that $\forall$ a $\in$ A determines a function $a: U \rightarrow V_a$.

B. Indiscernibility Relation

The starting point of RST is the indiscernibility relation [15], which is generated by information about objects. With every subset of attributes B of A there is an associated indiscernibility relation on U defined in Eq. (1).

$$IND\ (B) = \{\ (x,y) \in U \times U: f_a(x)=f_a(y)\ \} \tag{1}$$

The family of equivalence classes generated by IND(B) is called B-indiscernible and is expressed as U/IND(B) or U/B and is given by the Eq. (2)

$$U/IND(B) = \otimes\{U/IND(\{b\}) \mid b \in B\} \tag{2}$$

Let P and Q be two non empty finite sets then operation $\otimes$ is defined in Eq.(3)

$$P \otimes Q=\{X \cap Y \mid X \in P,\ Y \in Q,\ X \cap Y \neq \emptyset\} \tag{3}$$

C. Decision-Relative Discernibility Matrix

Let DS=(U,C,D) represents a Decision System[17], where U is the nonempty finite set of objects called the Universe and C is the finite nonempty set of conditional attributes and D is the finite nonempty set of decision attributes.

A Decision-relative discernibility matrix [17] is a symmetric matrix of order N, where N is the number of instances, and its entries are defined by Eq. (4)

$$D_{ij}=\begin{cases}\emptyset, if\ and\ only\ if\ d(xi) = d(xj)\\ dij, \qquad\qquad Otherwise\end{cases} \tag{4}$$

where $d_{ij} = \{c \in C: c(x_i) \neq c(x_j)\},\ for\ i,j = 1....,N$

A Discernibility function f_D is a Boolean valued function of m variables $b_1,\ldots,b_m$ corresponding to the m conditional attributes $c_1, \ldots, c_m$, can be defined as

$$f_D(b_1,...,b_m) = \wedge\{\vee D_{ij} \mid 1 \leq j \leq i \leq N,\ d_{ij} \neq \emptyset\} \tag{5}$$

D. Approximations

Let B$\subseteq$C and $D_a \subseteq$D then IND [B] represents the set of equivalence classes and is denoted by $[x]_B$. Let T be the target set such that T$\subseteq$U then the lower approximation of set T using the information in B is the set of objects that are inevitably belongs to the subset of interest, and upper approximation of set T is the set objects that possibly belong to the subset of interest. The equations for Lower approximation and Upper approximations for B are given by

$$\underline{B}T = \{\ t \mid [t]_B \subseteq T\} \tag{6}$$

$$\overline{B}\,T = \{\ t \mid [t]_B \cap T \neq \emptyset\} \tag{7}$$

The positive and negative regions are defined as

$$POS_B(D_a) = \cup_{T \in U/Da}\ \underline{B}T \tag{8}$$

$$NEG_B(D_a) = U - \cup_{T \in U/Da}\ \overline{B}T \tag{9}$$

E. Dependency of Attributes

Let P1 and P2 be two subsets of A. we say that P1 depends on P2, if and only if IND(P2)$\subseteq$ IND(P1) and is denoted as $P2 \underset{k}{\Rightarrow} P1$, where 0 $\leq$k$\leq$ 1 and is defined as

$$k = \frac{|POS\ P1(P2|)}{|U|} \tag{10}$$

where $POS_{P1}\ (P2) = \cup_{T \in U/P2}\ \underline{P1}\,T$

F. Reduct and Core

Reduct is a minimal subset of attributes that are essential and sufficient for the classification of objects of the universe and is defined as

$$Red\ (C_a) = \{R \subseteq C_a \mid \gamma_R(D_a) = \gamma_C(D_a), \forall S \subseteq R, \gamma_S(D_a) \neq \gamma_R(D_a)\} \tag{11}$$

There exists several reducts for a given dataset and the optimal set of reducts is given by

$$Red\ (C_a)_{min} = \{R \in Red \mid \forall R' \in Red, |R| \leq |R'\} \tag{12}$$

A Core is the most significant set of attributes and removal of a single element is not possible i.e., Core contains the set of all essential attributes.

$$Core\ (S) = \cap Red(S) \tag{13}$$

G. Quick Reduct Algorithm

QuickReduct algorithm [9] is a forward reduct generation algorithm. It starts the reduct computation process with an empty reduct set and recursively adds attributes one after one that result in the greatest increase in the rough set dependency metric, until a maximum possible value has been produced. Pseudo code for QuickReduct generation is given below.

QuickReduct (C_a, D_a)
Input: C_a, the set conditional attributes;
D_a, the set of decision attributes.
Output: RS, minimal subset of conditional attributes.

(1) RS ← { } (empty set)
(2) Do
(3) TS ← RS , $\forall x \in (C_a - RS)$
(4) If $\gamma\{RS \cup x\}$ $(D_a) > \gamma TS$ (D_a) then
(5) TS ←RS∪{x}
(6) RS←TS
(7) Until γRS $(D_a) = \gamma$ C_a (D_a)
(8) Return RS

The problem with Quick Reduct algorithm is, it may not scanning all attributes in the information systems and it is not practical for high dimensional datasets because of the increased processing times for computing the attribute dependencies.

H. Johnson's Algorithm

Johnson's algorithm [9] is a single reduct computation algorithm. The process of reduct generation starts with an empty set, RS. Iteratively, each conditional attribute in the discernibility matrix is evaluated based on a heuristic measure and the highest heuristic valued attribute is to be added to the RS and deletes the same from the original discernibility matrix. The algorithm ends when all the clauses are removed from the discernibility matrix. Pseudo code for Johnson's reduct generation is given below.

Johnson (C_a, f_D)
Input: C_a, the set of conditional attributes,
f_D is the Discernibility function.
Output: RS, The minimal reduct set

(1) RS← Ø; bestca=0;
(2) While (discernibility function, f_D is not empty)
(3) For each c ∈ C_a that appears in f_D
(4) h = heuristic (c)
(5) If (h > bestca) then
(6) bestca=h;
(7) bestAttribute ← c
(8) RS ← RS bestAttribute
(9) f_D ← removeClauses (f_D, bestAttribute)
(10) Return RS

The reduct generated by the Johnson's algorithm may not be optimal.

IV. Evolutionary Computation Algorithms

Evolutionary Computation algorithms[18] use mechanisms inspired by biological evolution process and natural inheritance. The popular EC algorithms are Genetic Algorithm, Ant Colony Optimization, Artificial Bee Colony Optimization and Particle Swarm Optimization.

A. Genetic Algorithm

Genetic Algorithm[19], developed by John Holland in 1975 is a global heuristic search technique used to obtain approximate solutions to optimization problems. GA's generate high quality solutions very quickly. Pseudo code for GA is given below.

Genetic Algorithm (DT, RS)
Input: DT, Decision Table
Ouput: RS, Reduct Set with minimum attributes

(1) Generate initial population (group of individuals)
(2) Evaluate fitness values of all individuals
(3) While (termination condition is not met) do
(4) Select parents with best fitness value
(5) Apply crossover to produce children
(6) Apply mutation on children
(7) Generate new population
(8) End while
(9) Return the individual with best fitness score as the RS.

The fitness function measures the quality of the represented solution. The definition of fitness function is problem dependent and choosing an appropriate fitness function is the hardest part of the GA.

The Classical, binary GA for reduct computation uses the fixed size solution representation i.e., bitmap is used to represent an individual. According to the definition of reduct, reduct is a minimal set of attributes. Bitmap is a binary string, its length is same as the total number of attributes and if a bit is set means the corresponding attribute of the decision table is in reduct. For example:

1011010001 → $\{A_1, A_3, A_4, A_6, A_{10}\}$ is the reduct.

To obtain the initial population, first step is to transform the given decision table into a distinction matrix.

Distinction matrix is a binary matrix DM, each column corresponds to one attribute and each row corresponds to one pair of different objects.

Let dm (i,(k,n)) be an element of distinction matrix, corresponding to the i^{th} attribute and pair (O_k, O_n) and N is the number of conditional attributes.

$$dm\ (i,(k,n)) = \begin{cases} 1, ai(Ok) \neq ai(On) \\ 0, ai(Ok) = ai(On) \end{cases} \quad (14)$$

$$dm\ (N+1,\ (k,n)) = \begin{cases} 1,\ di(Ok) = di(On) \\ 0, di(Ok) \neq di(On) \end{cases} \quad (15)$$

The initial population is generated from the distinction table, i.e., the row-wise entries of the distinction table form the initial population.

The fitness function of an individual represented by bitmap R is given by the eq. (16).

$$Fitness(R) = \frac{N-Lr}{N} + \frac{2Cr}{m(m-1)} \quad (16)$$

Where L_r denotes the number of 1's in the string R,
C_r denotes the number of object pairs discerned by the subset R and
m is the total number of instances.

The evolution process starts by selecting parents based on Roulette wheel selection algorithm and then generates the new population by applying genetic operators. The GA ends after a pre-determined number of generations.

The correctness of the GA lies in defining the fitness function. Selecting the appropriate representations, encodings, the positions of crossover points and mutation bits all together drives the GA to achieve optimal solutions.

B. Particle Swarm Optimization

PSO [6] is a introduced by Eberhart and Kennedy in 1995 as a new optimization technique inspired by sociological and biological behaviors of fish and birds. PSO is an EC technique, in which individuals communicate directly or indirectly with one another in the solution space to search for an optimal solution.

PSO starts its process by generating a population of solutions called particles and each particle in the population is considered as a point in a N-dimensional space represented by $X_i=(x_{i1},x_{i2},\ldots,x_{iN})$. The position and velocity of the particles is represented by $V_i=(v_{i1},v_{i2},\ldots,v_{iN})$ and the velocity of the particles are limited to V_{max}, maximum velocity. The acceleration of each particle towards the optimal solution is affected by its best previous position represented by pbest and the best known positions in the population are represented by 'gbest', the global best position. The movement of the particles in the search space is defined by eq.(17)

$$v_{id}^{t+1}=w*v_{id}^{t}+c_1*rnd1()*(p_{best}-x_{id}^{t})+c_2*rnd2()*(g_{best}-x_{id}^{t}) \quad (17)$$

$$x_{id}^{t+1}=x_{id}^{t}+v_{id}^{t} \quad (18)$$

Where v_{id}^{t+1} is the position and velocity of i^{th} particle in $t+1^{th}$ iteration, w is the inertia weight, c1 and c2 are constants called the accelerating terms that accelerates each particle toward the best positions and the functions rnd1() & rnd2() generates random values in between 0 & 1.

Pseudo code of Particle Swarm Optimization

PSO (Swarm, Pgbest)
Input: Swarm, the initial population
Output: Pgbest, the global best position is the optimal solution
(1) initialize the swarm
(2)Evaluate fitness of each particle
(3) while (terminate condition not met) do
(4) update individual and global bests
(5) update position and velocity of each particle
(6) end
(7) return the global best Pgbest as the optimal solution.

The classification error rate of the selected features is taken as the fitness function.

$$Fitness= error\ rate=\frac{FP+FN}{TP+TN+FP+FN} \quad (19)$$

Where TP stands for true positives and TF for true negatives,
FP stands for false positives and FN false negatives.

The parameters are set as follows:
Inertia weight, w=0.33,
Accelerating constants c_1= 0.34 and c_2=0.34,
Maximum velocity of the particles V_{max} is 6.0,
Initial population size is 50,
Set the maximum number of iterations to 100.

V. METHODOLOGY

The methodology of the proposed work is given in Fig.1

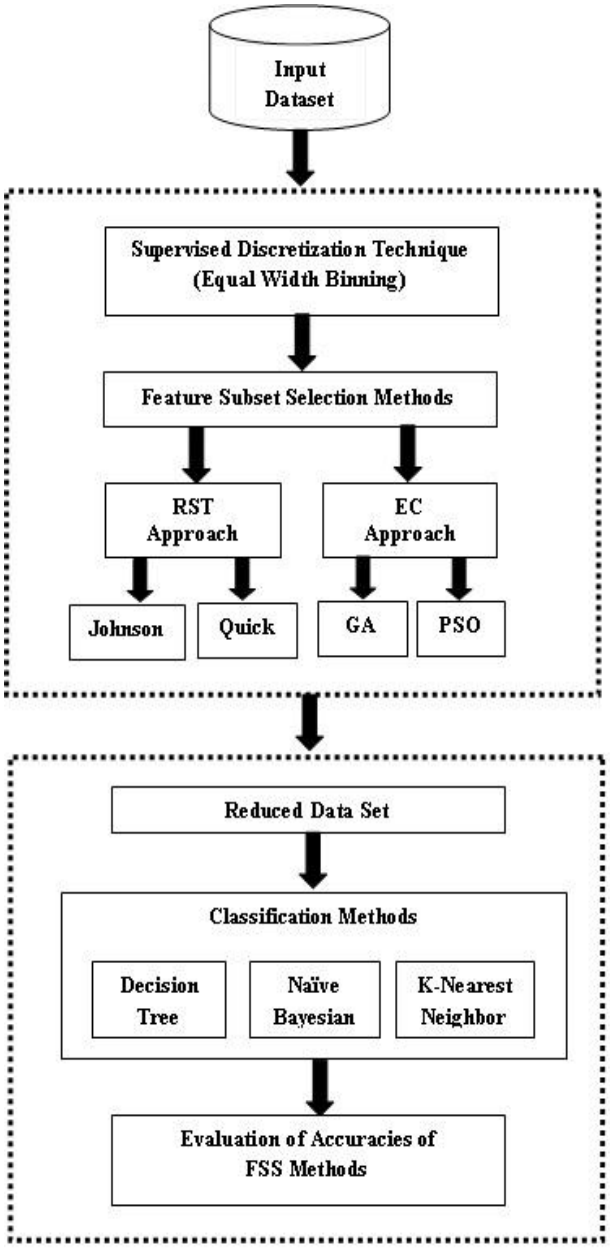

Fig. 1 Methodology of the proposed work

The proposed methodology is divided into three phases, first Discretization, second applying feature subset selection techniques and third phase is classification.

Phase 1: Discretization

Many real world datasets are consisting of continuous values i.e., too many possible values of an attribute for example, Age, Date etc. Discretization [20] is an important and becoming an essential pre-processing step in data mining algorithms because many of the data mining algorithms are proved to be efficient for discrete values. Discretization reduces the number of possible values by partitioning into intervals and finally improves the quality of data. For example, Age is a continuous valued attribute that takes possible values from 1 to 100, instead of taking the individual's age as a value , grouping of the values into three groups like Young Age, Middle Age, Old Age will enhance the readability and aids in classification process to get better results. Hence, discretization of dataset is the first step in the methodology.

Phase I1: Applying FSS Methods

During Phase II, apply FSS methods to find the relevant attributes from the given dataset using RST and EC approaches. FSS selects the significant features that contribute maximum to the classification. The Quickreduct and Johnson's reducer are the two reduct computation algorithms based on RST approach and the GA and PSO are the two popular efficient and effective EC algorithms for FSS. .

Phase II1: Classification

In phase III, reduce the dimensionality of the dataset by removing all irrelevant attributes from the dataset by keeping only the relevant attributes obtained by the above mentioned FSS methods and apply classification methods to the reduced dataset to evaluate the performances of the specified FSS methods. Classification [22] is one of the most common data mining techniques used to predict the target class of a new object on the basis of available information for a set of selected input attributes.

A. Decision Tree Classifier

Decision trees [23] are one of the most popular eager learning classification approaches. The Decision Tree (DT) is a tree like structure consisting of three types of nodes. The node with no incoming edges is called the root node and acts as the origin of the DT. The nodes having exactly one incoming edge and one or more outgoing edges are called internal nodes and the nodes with no outgoing edges are called leaves also known as terminal nodes or decision nodes. Each leaf node is assigned one class label. New objects are classified by traversing the tree starting from the root and down to a leaf, based on the outcome of the tests along the path and finally the label of the leaf will be considered as the target class of the unknown object. The advantage of DT classifier is very easy to understand, can handle both categorical and numerical data and performs well with large datasets.

B. K-Nearest Neighbor Classifier

K-NN [24] is a lazy learner, operates on the premises (data instances) and classifies new instances by relating the target class to the known instances class label according to some similarity or dissimilarity metric. K-NN is also known as Instance-based learner because it predicts the target class label of a new object based on the information taken from its 'K' nearest neighbors (i.e., K instances that are similar to the new object). One advantage of K-NN classifier is its learning time is very less and is easy to implement as the learning process is transparent. The major problem with this simple approach is very slow at query time i.e., cost of processing a new instance is high, lack of robustness and the high degree of local sensitivity is making K-NN classifiers highly vulnerable to noise in the training data.

C. Naïve Bayesian Classifier

Naïve Bayesian (NB) classifier [25] is based on the Bayes theorem and conditional probabilities. NB classifier is applicable only when the value of any specific feature in dataset is independent of all other features for a given target class. Naive Bayes classifiers can be trained efficiently with less computational time as they are based on probability concepts. The advantage of the Naive Bayesian classifier is that a very less amount of training data is sufficient for the computation of the conditional probabilities and to estimate the parameters that are necessary for classification.

VI. Experimental Analysis

For experimental analysis, the Thyroid [28] dataset is taken from UCI machine learning repository [29] consisting of 300 samples with no missing values. The dataset samples consist of 148 Hypothyroid, 41 Hyperthyroid and 111 negative patient's records. The dataset consists of 28 attributes, including 26 conditional attributes representing the symptoms and blood tests for the diagnosis of thyroid disease and two decision attributes (Class & Referral Source) representing the three types of thyroid disease and the suggested health centre for further proceedings. The samples consist of six continuous attributes and 20 categorical attributes. The six continuous attributes are preprocessed using Equal width binning technique of WEKA Tool's discretizer. As discretization makes learning process faster and more accurate, the accuracies of the classifiers are increased for the discretized dataset.

The accuracy measures of the three classifiers Decision Tree, K-Nearest Neighbor and Naïve Bayesian are evaluated using 10-fold cross validation test. The accuracies of the DT, K-NN and Naïve Bayesian classifiers on the discretized dataset have been increased from 95.47%, 75.72% and 87.24% to 96.70%, 95.88% and 96.29% respectively.

Fig.2 shows the affect of discretization on accuracies of three classifiers namely DT, K-NN and NB. After discretization the classification accuracy of K-NN increased nearly 20% and Naïve Bayesian's by 10%.

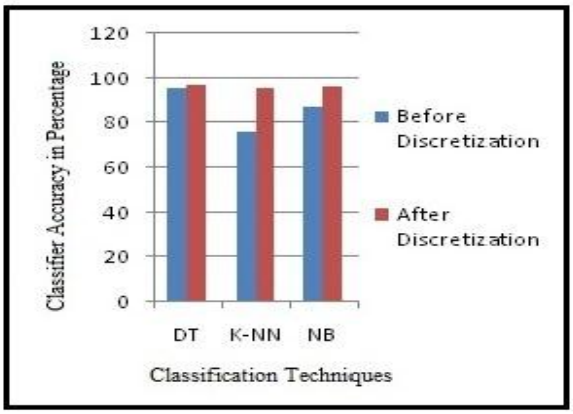

Fig. 2 Result of Discretization on Original Dataset

The EC based GA algorithm generated two reducts each with nine attributes; the feature subset that gives higher classification accuracies for all three classifiers is considered for comparing with other feature selection techniques. The performance of GA for the three classifiers is given in Table I.

TABLE I
PERFORMANCE OF GENETIC ALGORITHM

S.No	Size of Feature Subset	Decision Tree	K-Nearest Neighbor	Naïve Bayesian
1	9	97.53%	95.88%	97.53%
2	9	96.70%	95.88%	96.70%

The performance of PSO for different learning algorithms including the size of the feature subset and the observed classification accuracies is given in Table II.

TABLE II
PERFORMANCE OF PARTICLE SWARM OPTIMIZATION

Learning Algorithm	Size of Feature Subset	Decision Tree	K-Nearest Neighbor	Naïve Bayesian
K-NN	4	97.53%	97.94%	97.53%
DT	3	96.70%	96.29%	95.88%
NB	2	97.53%	97.53%	97.53%

The performance of PSO is significant when K-NN is taken as the learning algorithm. So, the feature subset that was generated when the K-NN algorithm was employed as the learning algorithm is considered as the optimal feature subset for comparing with other feature selection techniques.

The list of the minimal set of features obtained for EC based algorithms (PSO and GA) and RST based reduct computation algorithms (Quick reduct and Johnsons reducer) is given in Table III.

TABLE III
FEATURE SUBSET FOR VARIOUS FSS METHODS

Algorithm	No. of Attributes	Feature Subset
PSO	4	{Onthyroxine, TSH, T3, TT4}
Genetic Algorithm	9	{Sex, Onthyroxine, Sick, Tumor, TSH, T3, T131measurement, QueryHyperthyroid, TT4}
Quick Reduct	12	{Sex, Onthyroxine, Thyroid Surgery, Onantithyroxinemedication, Sick, TSH, T131measurement, QueryHypothyroid, QueryHyperthyroid,Psych,T3,TT4}
Johnsons Reducer	14	{Age,Sex,Onthyroxine,Queryonthyroxine, Onantithyroidmedication,Sick,Psych, T131measurement, QueryHyperthyroid, Tumor,TSH,T3, TT4,FTI}

The minimal set of features generated by the RST based approach is 12, but for the EC based approach the minimal set is reduced to 4, which is a great improvement.

The accuracy measures of DT, K-NN and Naïve Bayesian classifier for various FSS approaches are given in Table IV. The EC approach of feature subset selection techniques generated minimal subset of attributes while retaining the classification accuracy.

TABLE IV
CLASSIFICATION ACCURACIES FOR VARIOUS FSS APPROACHES

FSS Algorithm	No. of Attributes	Decision Tree	K-Nearest Neighbor	Naïve Bayesian
PSO	4	97.53%	97.94%	97.53%
GA	9	97.53%	95.88%	97.53%
Quick	12	97.53%	96.29%	97.53%
Johnsons	14	96.70%	95.88%	96.29%

The classification accuracies obtained on reduced dataset for the three classifiers DT, K-NN and Naïve Bayesian are shown in Fig. 3.

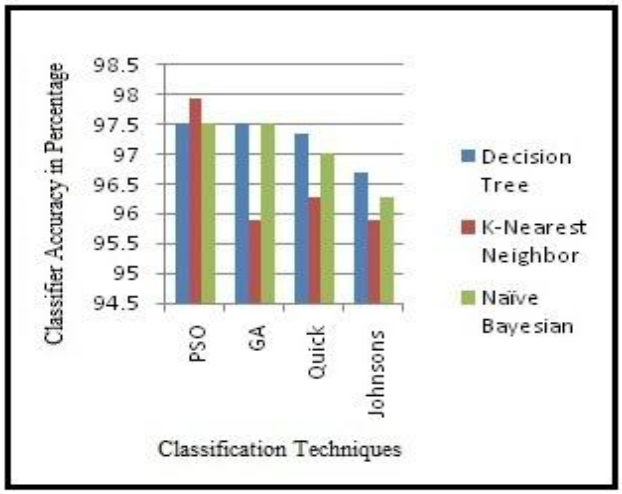

Fig. 3 Comparison of the performances of various Feature Subset Selection Methods

VII. Conclusion

Thyroid is one of the common diseases affecting many of the women during pregnancy and menopause. Thyroid disorder is often misdiagnosed and left untreated for long period requires immediate injection of thyroid hormone. Diagnosis of Thyroid disease is a process that includes various factors like blood tests, imaging tests and queries about the patient's health condition. So, from this list of symptoms and tests we need to identify a few that are most informative and helpful for the diagnosis of Thyroid. In this paper, to find the most significant and essential features that contribute maximum for the diagnosis of Thyroid disorder, various FSS methods have been applied to the Thyroid dataset. The two main RST based reduct computation algorithms; Johnson's and Quickreduct algorithm generated reducts with 14 and 12 attributes respectively. The two popular EC algorithms GA and PSO identified minimal subset of features with 9 and 4 attributes only. The performance of the three classifiers DT, K-NN and NB was observed to be increased for the reduced dataset with few attributes. Experimental results showed that the EC approach based PSO algorithm outperforms the GA and RST based algorithms, both in terms of minimal subset generation and increased accuracy of the classification.

This work can be further enhanced to diagnose the subclasses of hypothyroid i.e., Primary, Secondary or Compensated hypothyroid levels and can suggest any best health care centre as a Referral Source to the patients for further proceedings.

References

[1] L.Ladha and T.Deepa, "Feature Selection Methods and Algorithms," International Journal on Computer Science and Engineering (IJCSE), vol.3, no.5, pp. 1787-1797, May 2011.

[2] K.J.Cios, W.Pedrycz and R. Swiniarski, Data Mining Methods for Knowledge Discovery, Kluwer Academic Publishers. Poor, An Introduction to Signal Detection and Estimation. New York: Springer-Verlag, 1985, ch. 4, Nov 1998.

[3] R. Kohavi and G. H. John, "Wrappers for feature subset selection," Artificial Intelligence, vol. 97, pp. 273–324,1997. http://dx.doi.org/10.1016/S0004-3702(97)00043-X

[4] Z.Michalewicz, "Heuristic methods for Evolutionary computation techniques", Springer link journal of Heuristics, vol.1,pp.177-206,1996.

[5] J.Wr´oblewski, A.Skowron and L. Polkowski, "Genetic algorithms in decomposition and classification problem, Rough Sets in Knowledge Discovery 1", PhysicaVerlag, Heidelberg, vol.1, pp. 471–487,1998.

[6] J. Kennedy and R. Eberhart, "Particle swarm optimization," IEEE International Conference on Neural Networks,vol. 4, pp. 1942–1948,1995. http://dx.doi.org/10.1109/ICNN.1995.488968

[7] J.Starzyk, Dale E. Nelson and K.Sturtz, "Reduct Generation in Information System," Bulletin of international rough set society vol.3, pp.19-22, 1999.

[8] Y.Yao and Y.Zhao, "Discernibility matrix simplication for constructing attribute reducts", Information Sciences,179, no.7, pp. 867-882,2009. http://dx.doi.org/10.1016/j.ins.2008.11.020

[9] T.Chowdhury, A.Choudhary, P.Patwari and Saha, "Rules Mining for the Diagnosis of Breast Adenocarcinoma using Rough set theory," International Journal of Advanced Tech. & Engineering Research, l3, pp.14-21,2013.

[10] L. Y. Chuang, S. W. Tsai, and C. H. Yang, "Improved binary particle swarm optimization using catfish effect for feature selection," Expert Systems with Applications, vol. 38, pp. 12 699–12 707, 2011.

[11] B. Xue, M. Zhang, and W. Browne, "Novel initialization and updating mechanisms in PSO for feature selection in classification," in Applications of Evolutionary Computation, ser. Lecture Notes in Computer Science, vol. 7835, pp. 428–438,2013. http://dx.doi.org/10.1007/978-3-642-37192-9_43

[12] Z. Pawlak, "Rough Sets", International Journal of Computer and Information Science, (a seminal article), vol.11, pp. 341 – 356,1982.

[13] Z. Pawlak, "On rough dependency of attributes in information systems", Bulletin of the Polish Academy of Sciences, vol.33, pp. 551–599,1985.

[14] Z.Pawlak, "Rough Sets. Theoretical Aspects of Reasoning about Data," Kluwer Academic Publications, Netherlands,1995.

[15] Z.Pawlak, "Rough Set approach to knowledge-based decision support," European Journal of Operational Research, 99, pp. 48-57,1997. http://dx.doi.org/10.1016/S0377-2217(96)00382-7

[16] L.A.Zadeh, "Fuzzy sets",Information and control, vol.8,pp.338-353,1965. http://dx.doi.org/10.1016/S0019-9958(65)90241-X

[17] Son, C.S., "Decision-making model for early diagnosis of congestive heart failure using rough set and decision tree approaches", Journal of Biomedical Informatics, vol.45, pp.999-1008,2012. http://dx.doi.org/10.1016/j.jbi.2012.04.013

[18] T.Back, D.B.Fogel and Z.Michalewicz, "Handbook of Evolutionary Computation", IOP Publishing Ltd., 1997. http://dx.doi.org/10.1887/0750308958

[19] J.Wr´oblewski, A.Skowron and L.Polkowski, "Genetic algorithms in decomposition and classification problem," Rough Sets in Knowledge Discovery 1, PhysicaVerlag, Heidelberg, 1, pp. 471–487,1998.

[20] H.Liu, F.Hussain,C.L.Tan and M.Dash, "Discretization: An Enabling Technique, Data mining and Knowledge Discovery," Kluwer Academic Publications, Netherlands, 6, pp. 393-423,2002.

[21] X.Bing, A.K.Qin and M.Zhang, "Particle swarm optimization for feature selection in classification", IEEE congress on Evolutionary Computation, vol.6, pp.3119-3126, 2014.

[22] J.G.Bazan, "Rough set Algorithms in Classification Problem," Springer Rough set Methods and Applications, Physica-Verlag2,2000.

[23] L.Rokach and O. Maimen, "Data Mining and Knowledge Discovery Handbook", 9,pp.165-192,2010.

[24] P.Cunningham and S.J.Delany "K-Nearest Neighbor Classifiers," Multiple Classifier Systems, Technical Report UCD-CSI-2007, 4,2007.

[25] K.P.Murphy "Naïve Bayes Classifiers," University of British Columbia,2006.

[26] K.Thangavel and A. Pethalakshmi, "Dimensionality reduction based on rough set theory," Applied Soft Computing, Elsevier, 9, 1-12.2009.

[27] N. Ibrahim, T.Hamza, and E.Radwan, "An Evolutionary Machine Learning Algorithm for Classifying Thyroid Diseases Diagnoses," Egyptian Computer Science Journal, 35, pp.73-86,2011.

[28] M.P.Vanderpump, "The epidemiology of thyroid disease," British medical bulletin 99,pp.39-51,2010. http://dx.doi.org/10.1093/bmb/ldr030

[29] https://archive.ics.uci.edu/ml/datasets/Thyroid+Disease

S.Surekha is currently working towards her Ph.D in JNTUK-University College of Engineering, Kakinada. She is working as Assistant Professor in the department of Computer Science and Engineering, JNTUK-University College of Engineering, Vizianagaram. Her research includes Data Mining and Machine Learning.

Mrs.Surekha is a Life member of Computer Society of India.

G.Jaya Suma is Head of the Department, Department of Information Technology, JNTUK - University College of Engineering, Vizianagaram. She received her Ph.D from Andhra University in 2011. Her current research includes Data Mining, Soft computing and Mobile Computing.

Dr.JayaSuma is a Life member of ISTE, a member of IEEE and CSI.