Soft Comput
DOI 10.1007/s00500-014-1581-5

METHODOLOGIES AND APPLICATION

Thyroid disease diagnosis via hybrid architecture composing rough data sets theory and machine learning algorithms

V. Prasad · T. Srinivasa Rao · M. Surendra Prasad Babu

Abstract This work is a hybrid architecture design furnished successfully using artificial intelligence techniques, rough data sets theory and machine learning algorithms. The purpose of this work is to bring the spotless and smart approach in identifying the thyroid disease in a human. There are several mechanisms implemented on thyroid data sets which produced astonished outcomes, but the data considered for the thyroid disease diagnoses (TDD) is inconsistent, redundant and consists of missing attribute values as per my literature survey. The proposed work is to construct an expert advisory system of hybrid architecture, which is to determine the optimistic disease growth because of the thyroid gland. A string matching system (SMS) was at the outset developed, which can predict the actual TDD based on the knowledge available. If the SMS fails, an individual approaches using artificial bee colony optimization and particle swarm optimization are developed to achieve the accuracy of results appreciating the measure values as 65 and 93 %, but the results obtained using the above said approaches are calculated using some missed attribute values which are not included in the knowledge likewise left blank and hence the proposed work continues to first generate the missing attribute value in the knowledge by using rough data sets theory and the obtained data (missed attribute values) is given to predict optimistic disease along with its prevention and its curing methods. However, the data generated cannot predict the optimal disease and hence it is proposed to use a machine learning algorithm so that, obtained result is hygienic. The knowledge for implementation of this work is gathered from Intelligent System Laboratory of K.N. Toosi University of Technology, Imam Khomeini Hospital. A questionnaire form is developed for providing an interface for user so they can contribute the data.

Keywords Machine learning algorithms · Rough data sets · Medical diagnose · Thyroid disease expert advisory system · Hybrid architecture · Optimistic disease

1 Introduction

Thyroid disease (TD) is a study of Endocrinology, which at most seems to be an element in Science of Medicine. Thyroid is one of the most common diseases that is frequently misunderstood and misdiagnosed (Azar et al. 2012; Keles 2008). Based on the thyroid conditions the diseases which are obtained by the thyroid gland are classified as seven categories i.e., hyperthyroid, hypothyroid, binding protein, replacement therapy, anti-thyroid therapy, non-thyroid illness (general fever) and miscellaneous features. There are 28 pristine attributes which are considered for tracing out the specimens of TD, but considering this geographical world the thyroid gland effect will be different all around.

Nowadays, by developing technology and information in medical sciences, the computer science professionals are capable of providing expert advisory system (EAS)

Communicated by V. Loia.

V. Prasad (✉)
Department of CSE, Raghu Institute of Technology, Visakhapatnam, Andhra Pradesh, India
e-mail: prasad_v542@yahoo.co.in

T. S. Rao
Department of CSE, Gitam Institute of Technology, Gitam University, Visakhapatnam, Andhra Pradesh, India

M. S. P. Babu
Department of CS&SE, AU College of Engineering (Autonomous), Andhra University, Visakhapatnam, Andhra Pradesh, India

Published online: 13 January 2015

(Azar et al. 2012) to diagnose different kinds of diseases with high accuracy. The medical professionals are made to use these systems due to some developed errors during general diagnosis process. Disease diagnosis operations using EAS are performed based on sets of disease symptoms. These systems are based on artificial intelligence which helps the physician to minimize the costs and time in effective diagnoses. Among these, artificial neural network is a family of artificial intelligence, the researchers could reach to the big success using them in diagnosing diseases such as diabetes, heart disease, liver diseases and so on.

Inspired with the above said, we proposed a method of invoking machine learning algorithm (MLA) into this proposed work of thyroid disease diagnoses (TDD). MLA has advantages such as self-learning, high parallelism, speed and error tolerance against noise. The readiness of reuse feature in MLA made us to choose it as an option to develop this proposed work.

In this paper rough data sets (RDS) theory tries to classify TD in the presence of missing bases and build the modified similarity relations (MSR) (Radawan and Assiri 2013) that is dependent on the number of missing bases with respect to the number of the whole defined attributes for each rule. The thyroid related attributes are converted for suitable representation of rough set analysis. This is done in two levels i.e., discretizing and later a matrix is generated with rows representing similarity score of attributes and columns representing the definition of the attribute for individual rule set. The discernibility matrix produces new decision rules based on missing attributes, applying the procedure of MSR. Java Server Pages (JSP) is used in programming due to its flexibility. This made the output levels reach high accuracy, which in turn made MLA performance the best in short time.

In Sect. 2, we discuss several papers regarding TDD. In Sect. 3, a new approach to TDD is represented by proposing a method which is a combination of RDS and MLA. In Sect. 4, we made algorithms in detail followed by event flow diagram of the hybrid architecture. In Sect. 5, the results are discussed and last in Sect. 6, conclusion is made with all the advantages and disadvantages.

2 Literature survey

According to the previous research, various methods including multi-layer perception with back-propagation method (MLP), radial basis function (RBF) and adaptive conic section function neural network (ACSFNN) are used to help diagnosis of thyroid disease; their classification accuracies are separately 81 to 88 %. Five methods namely linear discriminate analysis (LDA), C4.5 with default learning parameters (C4.5-1), C4.5 with parameter c equal to 5 (C4.5-2), C4.5 with parameter c equal to 95 (C4.5-3) and DIMLP with two hidden layers (DIMLP) performed classification and the output accuracy levels are in between 89 to 94 %. Moreover, an accuracy of 81 % was obtained with the application of artificial immune recognition system (AIRS). In addition to the above, ESTDD (expert system for thyroid disease diagnosis) produced the accuracy levels at 95.33 %. Finally, swarm optimization optimized support vector machines with Fisher score (FS-PSO-SVM) CAD system for thyroid disease, and the average accuracy of 97.49 % was achieved (Gharehchopogh et al. 2013).

As a result, all MLA which are in use depend on different kinds of heuristics. The resultant recodes contain missing values, which leads to misclassifying of the diseases. Also, a continuous dataset due to the measuring range is discovered. Thus, a good and effective tool to deal with vagueness, missing and uncertainty of information is needed. Rough sets theory deals with the classification of data tables and focusses on structural relationships in data sets. Rough sets theory constitutes a framework for inducing minimal decision rules (Jacqulin Margre et al. 2012; El-Sayed El-Metwally et al. 2010), these rules in turn can be used to perform a classification task. The main goal of the rough set analysis is to search large databases for meaningful decision rules and finally acquire new knowledge. Rough sets (Peters et al. 2102; El-Sayed El-Metwally et al. 2010; Polkowski and Skowron 1998; Pawlak 1991) have been successfully applied in many different fields, particularly in the medical field (Akbas et al. 2013). A rough set investigates structural relationship in data rather than probability distribution and produces decision table (Jacqulin Margre et al. 2012) rather than trees.

Backdrops in the above said works

- Out of 30 features of TDD, only technical and repeated symptoms are considered for disease determination.
- Comparison is done for various algorithms implemented, which does not support same TDD with similar attributes.
- Applying more algorithms increases the efficiency for some cluster of data but it is not valid for a database which holds large knowledge base.
- Classification of all TDDs is not performed, generally all the works stopped at hypo, hyper and illness.
- Testing sensitivity of algorithm varies from one algorithm to another algorithm.
- Missing values disqualifying the original TDD.

3 Proposed system

The proposed system consists of major 12 attributes which are used for identification of the 7 categories of TD. The

work is a hybrid architecture (Prasad et al. 2013) which is composed of rough data sets theory, machine learning algorithms (Li and Zhou 2006; Guvenir et al. 1998) and expert advisory system (Azar et al. 2012). The system developed is used for offline analysis and optimistic approach on TDD. Originally the system is developed by using a string matching system (SMS) (Crochemore et al. 1994), when the SMS fails to identify the relevant data in the knowledge base, then automatically the system invokes to get the optimistic disease by using two individual methods (i) artificial bee colony (ABC) optimization (Karaboga and Akay 2009) (ii) particle swarm optimization (PSO) (Karaboga and Basturk 2007). The method of ABC on TDD produced a result of maximum range 73 %, where as the PSO on TDD produced 98 %, but observed that there are some missing attribute values which are to be calculated to give perfect optimistic solution. Hence, the system is further enhanced with rule based system (RBS) in order to provide the disease name with its prevention and curing methods (Prasad et al. 2014a). Here, we obtain the diagnostic methods for different symptoms entered by the user dynamically (Prasad et al. 2014b). If the data entered by the user is sufficient and if it matches the knowledge base (KB), then the proposed system displays the actual disease with which the human is suffering, or else it displays the dialogue box stating that the knowledge is insufficient. For the purpose of calculating missing attribute values, rough data sets (RDS) theory been introduced, so that the nearby related diseases of TD can be determined by the matrix oriented method of RDS, based on the raised values, if the exact disease is unavailable in the knowledge base then the HA system identically shows the probabilistic disease with which the human is suffering from. The particle swarm optimization technique is the combination of both ID3 and decision making algorithms that provides the actual disease (Akbas et al. 2013) by considering the values generated by the RDS, if the raised values satisfies the grouping methodology i.e., (swarming) then PSO is responsible for providing the optimized disease which could be predicted.

4 System design

In the development of this offline web portal we had implemented

(a) String matching system with individual ABC and PSO algorithms.

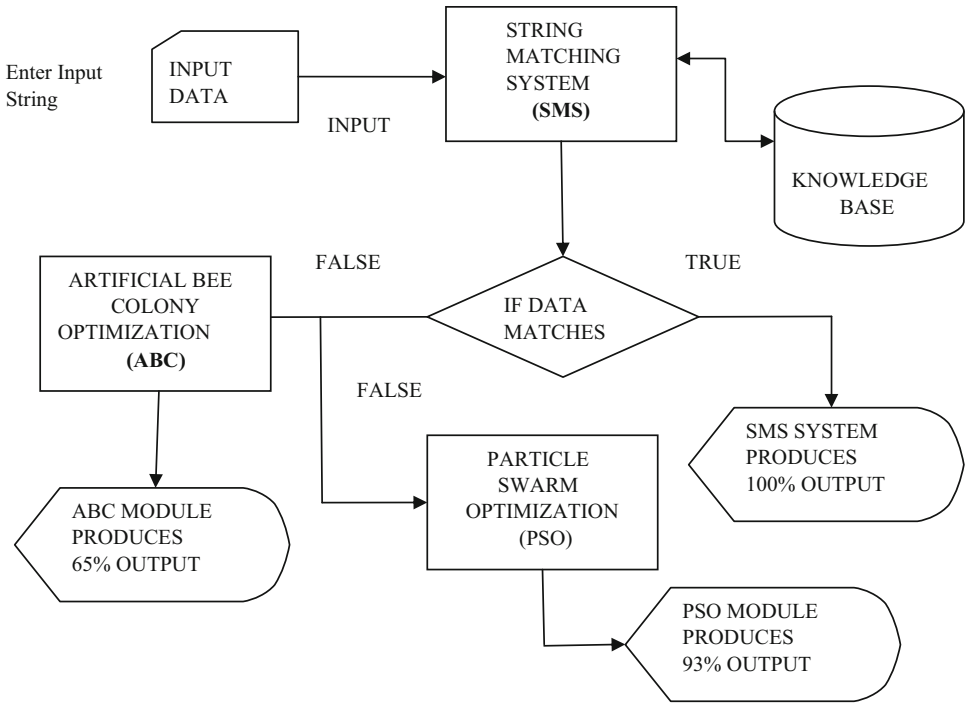

Fig. 1 The above diagram shows, that the input is given in a string format to the developed string matching system (SMS), if the 12 symptoms data entered by the user in the questionnaire form exactly fires the data in the knowledge base then, automatically the SMS is successful and it provides the information regarding the TDD along with its prevention and curing methodology, otherwise the entered input string is traversed to ABC optimization method which produces an approximation measure rate at 65 % and similarly the entered input string without any modification is given to PS optimization which produces 93 % output

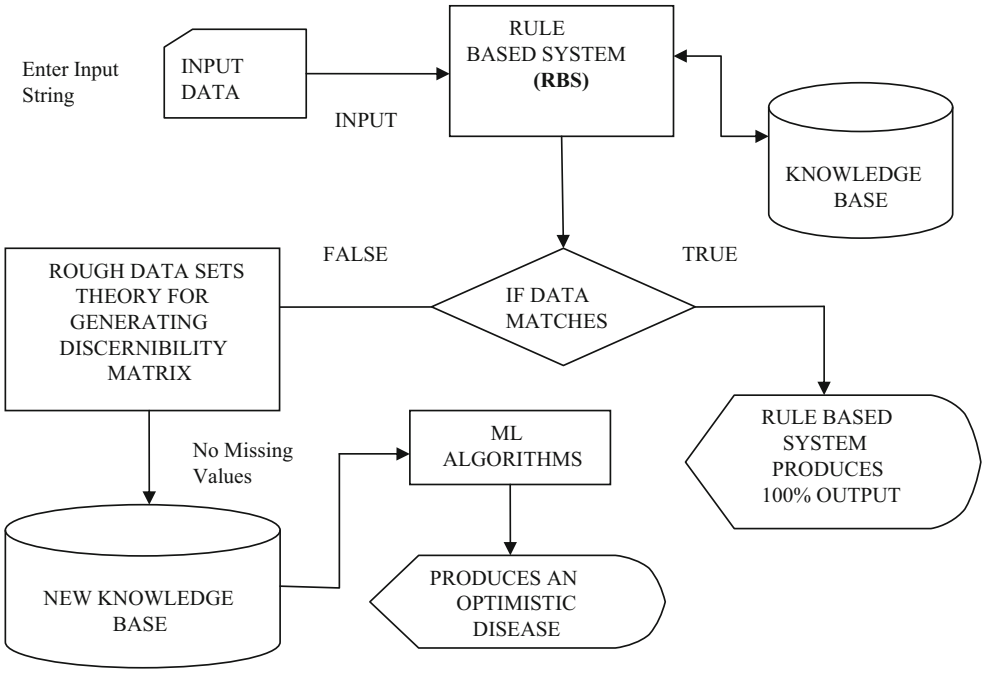

Fig. 2 The above diagram shows, that the input is given in a string format to the developed rule bases system (RBS), if the 12 symptoms data entered by the user in the questionnaire form, if exactly fires the data in the knowledge base then, automatically the RBS is successful and the information regarding the TDD along with its prevention and curing methodology is displayed, otherwise the entered input string is traversed to rough data sets theory, where RDS theory is designed in such a way that, a discernibility matrix which is used to trace the local max and min values for the former missed attribute value is produced to develop new knowledge base with a value approximation and now the new knowledge base obtained in given to particle swarm optimization which produces 98 % relevant measure rate which is better when compared

(b) Rule based system along with rough data sets generation and PSO algorithm.

Event flow diagram

See Figs. 1 and 2.

4.1 Rough set theory

Rough set theory (RST) is used for grouping thyroid attributes in the presence of missing attribute values and build modified similarity relation (MSR) that is dependent on the number of missing values with respect to number of whole defined attributes for each rule. Rough data sets deal with vague, missing and uncertainty values which intern provides a new knowledge which is worth useful.

RST is a good tool for data analysis. This theory is an observation to find out the similar relation (SR) with the existing available information. Tracing the similar objects by having the missing values is impossible. So, in order to work out with incomplete information system (IIS) there should be an extension to SR.

Data set is called an IS which is a data table, whose columns are labelled by attributes, rows are labelled by objects or cases, and the entire of the table are the attribute values. Formally, IS = (U, ATT), where U and ATT are nonempty finite sets called "the universe" and "the set of attributes," respectively. Every attribute, "A" belongs to ATT, has a set of its values V_a called the "domain of a". If V_a contains missing values for at least one attribute, then IS is called an IIS, otherwise it is complete. Any information table defines a function ξ that maps the directly U and ATT into the set of all values assigned to each attribute.

The example of IIS in Table 1 where set of objects in the universe corresponding to set of rules and set of attributes corresponding to set of values inside each new instance. The values of attributes are corresponding to the values of bases inside each object such as the value of Rule1 at S1is defined by ξ (Rule1, S1) = 0.

The concept of similarity relation (SR) is a concept in RST which is used to distinguish objects described by a set of attributes in complete information system (IS). Each subset A of ATT defines an SR as follows:

$$\mathrm{SR}(A) = \{(x, y) \in U \times U : \xi(x, a) = \xi(y, a) \quad \forall a \in A, A \subset \mathrm{ATT}\} \tag{1}$$

In, the above Eq. (1) SR(A) is the similarity relation or the equivalence relation which consists x as objects

Table 1 The incomplete information system gives data of missing attributes at following mapping as ξ (Rule1, S7) = *, ξ (Rule2, S2) = *, ξ (Rule3, S6) = *

Entered data	S1	S2	S3	S4	S5	S6	S7	S8	S9	S11	S11	S12
Rule1	0	1	1	1	0	1	*	0	0	0	1	0
Rule 2	1	*	0	0	0	1	0	1	0	0	0	*
Rule 3	0	1	1	0	0	*	0	0	1	1	0	0

Table 2 The partial information system which consists of only one missing value which can be traced through MSR and before that by using SR we obtained the missing values for the said ξ (Rule1, S7) = * as 0, ξ (Rule2, S2) = * as 1, ξ (Rule3, S6) = * as 1

Entered data	S1	S2	S3	S4	S5	S6	S7	S8	S9	S11	S11	S12
Rule1	0	1	1	1	0	1	***1**	0	0	0	1	0
Rule 2	1	***1**	0	0	0	1	0	1	0	0	0	*
Rule 3	0	1	1	0	0	***1**	0	0	1	1	0	0

and y as attributes along with a specified attribute value "a" $\subset A \subset$ ATT.

If the relations are similar then the Missing Values can be traced or else finding the SR for the missing values we can deduct a new knowledge, but the new knowledge arrived will be in two fundamental concepts (a) CORE (b) REDUCT.

REDUCT is a knowledge which can be used for finding the basic occurring of the value and CORE produces the most efficient value which is a certain sense and hence the CORE value defined as

$$\mathrm{SR}(A) = \mathrm{SR}(A - \{a\}) \tag{2}$$

The CORE is possible, only when there is no obligation for the attribute value "a" to be removed from its attribute list. If the obligation occurs then, the set of attributes which are non occurrence of "A" are considered to be "B".

$$\mathrm{SR}(B) = \mathrm{SR}(A) \tag{3}$$

and A can have any number of REDUCTS.

$$\therefore \mathrm{CORE}(A) = \cap \mathrm{REDUCT}(A) \tag{4}$$

Many researchers have proposed to represent knowledge in a form of matrix. This representation is used for computing CORE and REDUCT of knowledge.

Let NIS = (U, ATT) be a knowledge representation system with $U = \{x_1, x_2, x_3, x_4, x_5, \ldots, x_n\}$ by the matrix of NIS (new information system) denoted $M(k)$, which means a $n \times n$ matrix can be defined as

$$C_{ij} = \{a \in A : \xi(x_i, a) \neq \xi(x_j, a)\} \quad \text{for } 1,\ j = 1, 2, \ldots, n \tag{5}$$

Thus, the entry C_{ij} is the set of all attributes which discern objects x_i and x_j and the core can be defined as

$$\mathrm{CORE}(A) = \{a \in A : C_{ij} = (a) \text{ for some } i, j\} \tag{6}$$

and based on the CORE value and from its equivalent classes of the core relationships, we identify the REDUCT value based on Eq. (4).

The SR approach is useful to trace the missing value when there is an equivalent relations among the clusters generated but if there is no equivalent approach then the approach is not easy to calculate the CORE and REDUCT values because the values may vary.

Based on the equivalent relation values obtained generate a CUT-OFF (CO) points to obtain the CORE and REDUCT value without any disqualifying values as a content value (Table 2).

The RST approach used here is modified to deal with IIS, where IIS = (U, ATT) where U and ATT are non empty finite sets. Similarity relation SR(A) denotes a binary relation between objects that are possibly similar in terms of values of attributes and in case of missing values the modified relation is:

$$\mathrm{SR}(A) = \{(x, y) \in U \times U, a \in A, \xi(x, a) = \xi(y, a) \text{ or } \xi(x, a) = * \text{ or } \xi(y, a) = *\} \tag{7}$$

$$\mathrm{SR}(x) = \{y \in U : (x, y) \in \mathrm{SR}(A), A \in \mathrm{ATT}\} \tag{8}$$

The modified similarity relation (MSR) can be defined as:

- $(x, x) \in \mathrm{MSR}(A)$ where $A \subset \mathrm{ATT}\ \forall x \in U$.
- $(x, y) \in \mathrm{MSR}(A)$ where $A \subset \mathrm{ATT}$, $N = |A| \geq 2$ if and only if
  - $(\xi_x)^a = (\xi_y)^b, \forall a \in A$, where $(\xi_x)^a, (\xi_y)^b$ are defined as values
  - $\mathrm{EP}(x, y) \geq N/2$, if N is even
  - $\mathrm{EP}(x, y) \geq (N + 1)/2$, if N is odd

where $\mathrm{EP}(x, y) = |((\xi_x)^a, (\xi_y)^b)|$ for all $a \in A$, $A \subset \mathrm{ATT}$ is the number of equal pairs for the attribute "a" for all $a \in A$

Table 3 The complete information system consists of traced missing values by MSR; we obtained the missing value for the said ξ (Rule2, S12) = * as 0 instead of * by calculating the value as EP = $|((\xi_x)^a, (\xi_y)^b)|$

Entered data	S1	S2	S3	S4	S5	S6	S7	S8	S9	S11	S11	S12
Rule1	0	1	1	1	0	1	***1**	0	0	0	1	0
Rule2	1	***1**	0	0	0	1	0	1	0	0	0	*
Rule3	0	1	1	0	0	***1**	0	0	1	1	0	0

for the objects "x" and "y" respectively and $|((\xi_x)^a, (\xi_y)^b)|$ is the missed value obtained (Table 3).

4.2 Pseudo code of the ABC optimization implemented

Initialize ::

// Bees – Symptoms Entered
// Food Sources – Desired Disease
// Hive – Search Space

Procedure ::

Employed Bees – pbest, gives the accurate disease
This means ,75% best value is obtained
Onlookers – gbest , which gives u the neighbor disease
This means , 40% best value is obtained
Scouts – lbest, worst, where there are no matching symptoms in base
Exactly 0%

**This ABC algorithm is a technique in swarming algorithms and it is also used for solving discrete and constraint oriented problems (Radawan and Assiri 2013).

4.3 Pseudo code of the PSO procedure implemented

Step 1: Initialization of particles with random position and vectors.

Input the symptoms and convert them in a string level.

Step 2: For each particle evaluate its fitness.

Divide into groups by symptom wise and evaluate its fitness.

Step 3: Choose the Best Solution

if solution is found then carry on
if fails go to the neighbour solution and carry on
With the next level of solution

Step 4: Update the values for more optimized information

if any further information needed update it.

Step 5: Stop by giving best Solution.

Or if everything is clear, stop the process and display the output at level manner.

4.4 Functional requirements

The following are the functional requirements of this system:

4.4.1 Inputs: the system consists of both static, dynamic information (through JSP Interface) related to

- About thyroid diseases
- Thyroid disease classification
- Common thyroid disease symptoms
- Investigations
- Curing methodologies
- Preventions.

4.4.2 Outputs: the outputs of the system will be

- Information disease
- Description about the disease
- Diagnosis
- Healthy advice.

4.4.3 Store the information collected through inputs is stored as a knowledge base that serves as a repository for quick processing and future retrieval.

4.4.4 Computations Various computations are to be performed while evaluating and generating reports based on the requirement and information collected. These are nothing but forward chaining and backward chaining etc.

4.5 Non-functional requirements

4.5.1 Platform Windows platform that is equipped with Java Server Pages and Net Beans.

4.5.2 Technology to be used JAVA (JSP Front End) and MYSQL (Rear End).

4.5.3 Development process to be used. This is a web enabled application developed using JSP and Datasets. So as to ensure the quality of the software, all software engineering concepts, including test cases are implemented. This depicts the client server architecture and forms a well set layout.

4.6 Knowledge base

Knowledge base is constructed using 8 patients data of Andhra Pradesh with 57 rules and 390 patients data con-

Table 4 After considering the entered symptoms, the disease can be predicted from this sample database, like wise it consists of more than 2,000+ tuples for identification of the TD

Symptoms	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12	Diagnose	Prevention or referral source
Rule set-1	0	1	1	1	0	1	0	0	0	0	1	0	Thyroid is negative	SVHC (or) feel free, but a regular checkups on thyroxin is necessary
Rule set-2	1	0	0	0	0	1	0	1	0	0	0	0	Hyperthyroid	SVI (or) consult a near by doctor, since already thyroxin levels are more
Rule set-3	0	1	1	0	0	0	0	0	1	1	0	0	Thyroid primary, antithyroid therapy	SVHD (or) antithyroid medication is required and suggested to use antithyroid drugs
Rule set-4	0	0	0	0	1	1	0	1	0	0	1	0	Missing menstrual cycles and drowsiness and swellings	STMW (or) eat food which gains proteins in it and reduce the swellings by drinking less water and try to have water which is containing minerals in it
Rule set-5	0	1	0	0	0	0	1	1	1	1	1	0	Goitre	Others (or) swellings to be reduced and suggested antic radioactive ionized treatment

sisting 549 rules obtained from the Intelligent System Laboratory of K.N.Toosi University of Technology from Imam Khomeini hospital (Li and Zhou 2006).

Symptoms

S1 = Recognition of thyroxin S2 = Query on thyroxin
S3 = Anti thyroid medication S4 = Sick/weakness/ drowsiness
S5 = Pregnancy problems S6 = Thyroid surgery
S7 = I131 treatment S8 = Hypothyroid query
S9 = Hyperthyroid query S10 = Lithium
S11 = Goitre S12 = Tumor
Indication in data sets 1-Yes 0-No

Rule 1
If symptoms
S1 = 0, S2 = 1, S3 = 1, S4 = 1, S5 = 0, S6 = 1, S7 = 0, S8 = 0, S9 = 0, S10 = 0, S11 = 1, S12 = 0
Diagnose: thyroid is negative
Prevention or curing method: feel free, but a regular checkups on thyroxin is necessary.

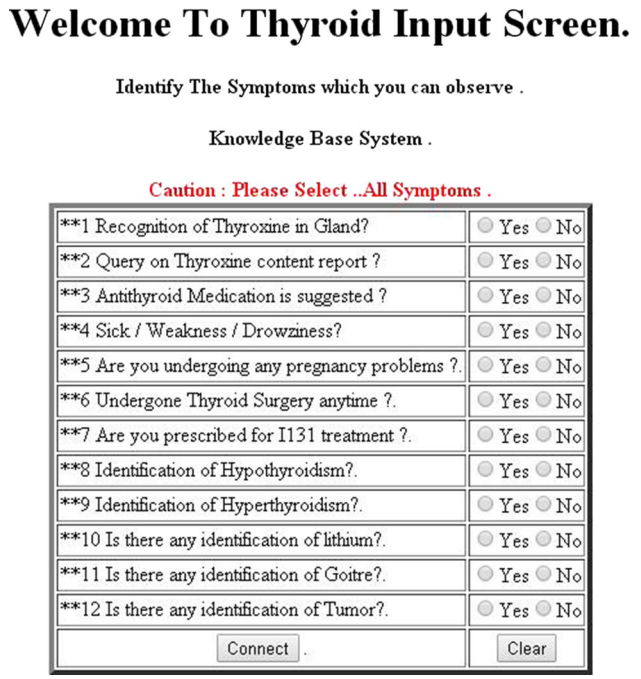

Fig. 3 This screen shot represents the JSP page developed as front end to serve as an interface between the sophisticated user and the knowledge base. Using this interface, the user enters the data which he identified or traced and connects with the KB

Rule 2
If symptoms
S1 = 1, S2 = 0, S3 = 0, S4 = 0, S5 = 0, S6 = 1, S7 = 0, S8 = 1, S9 = 0, S10 = 0, S11 = 0, S12 = 0
Diagnose: hyperthyroid
Prevention or curing method: referral source is SVI (or) consult a near by doctor,
Since already thyroxin levels are more.

Rule 3
If symptoms
S1 = 0, S2 = 1, S3 = 1, S4 = 0, S5 = 0, S6 = 0, S7 = 0, S8 = 0, S9 = 1, S10 = 1, S11 = 0, S12 = 0
Diagnose: thyroid primary, antithyroid therapy
Prevention or curing method: antithyroid medication is required and suggested to use antithyroid drugs.

Rule 4
If symptoms
S1 = 0, S2 = 0, S3 = 0, S4 = 0, S5 = 1, S6 = 1, S7 = 0, S8 = 1, S9 = 0, S10 = 0, S11 = 1, S12 = 0
Diagnose: missing menstrual cycles and drowsiness and swellings
Prevention or curing method: eat food which gains proteins in it and reduce the swellings by drinking less water and try to have water which is containing minerals in it.

Rule 5
If symptoms
S1 = 0, S2 = 1, S3 = 0, S4 = 0, S5 = 0, S6 = 0, S7 = 1, S8 = 1, S9 = 1, S10 = 1, S11 = 1, S12 = 0
Diagnose: Goitre
Prevention or curing method: swellings to be reduced and suggested antic radioactive ionized treatment.

Rule 6
If symptoms
S1 = 1, S2 = 1, S3 = 1, S4 = 1, S5 = 1, S6 = 1, S7 = 1, S8 = 0, S9 = 0, S10 = 1, S11 = 0, S12 = 1
Diagnose: T3 toxic
Prevention or curing method: thyroxin levels are more and requested to compel down by removing the goitre if observed and to reduce the thyroxin levels by daily checkups radio iodine treatment.

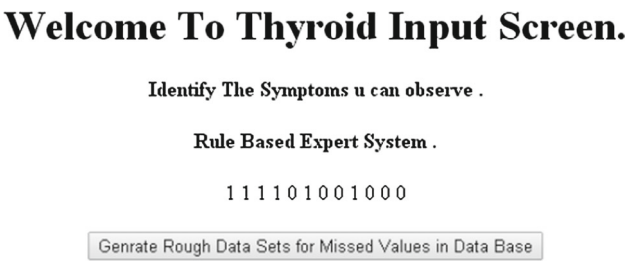

Fig. 4 Failure in SMS and RBS would leave the generated string displayed for the user information, the same string is used to process with RDS to generate new data which are nothing but the missing values, which can predict the optimized TDD. These missing values are generated based on their higher and lower cut off values for a attribute

The datasets mainly consists of 12 symptoms with which we can predict the existence of disease obtained through thyroid gland (Thyroid disease records. https://archive.ics.uci.edu/ml/datasets/Thyroid+Disease) (Table 4).

5 Results

CASE 1 :: SMS generates 100 % output.

If symptoms in the knowledge base 100 % matches, then the process of identifying the disease and providing the pre-

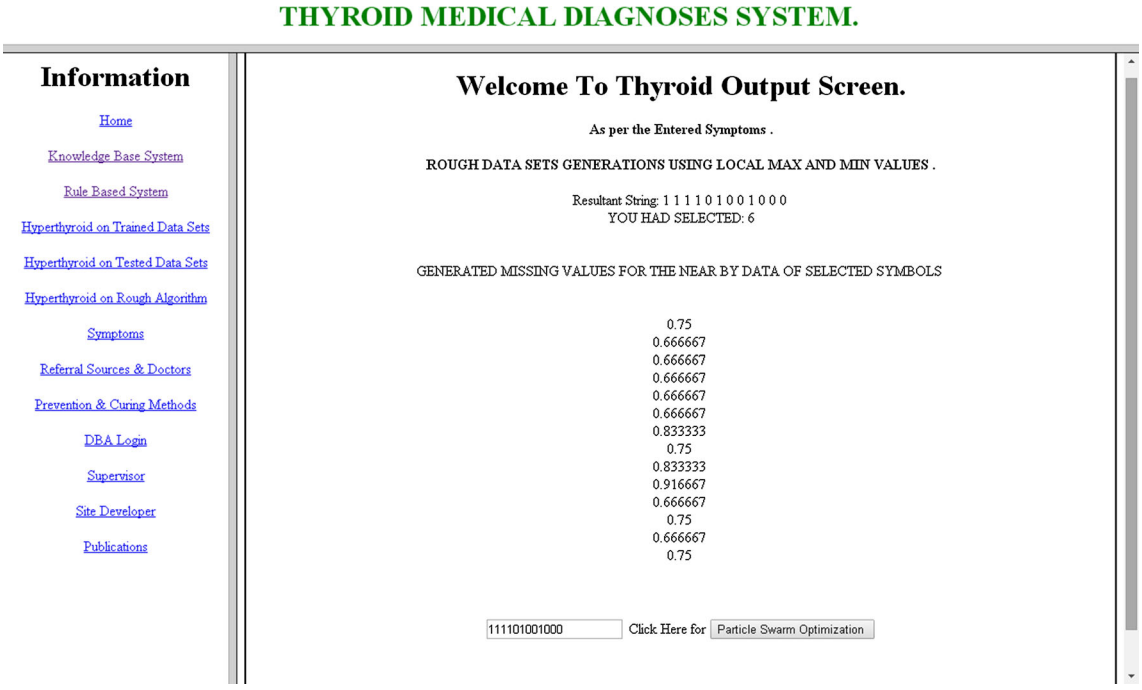

Fig. 5 Screen shot shows the complete Home page of the module designed with SMS (knowledge base system), rule based system for generated new databases with nearer values by RDS. This show the nearer values which can be raised in the levels of 0.666667 to 0.833333. The input string which is entered by the user along with the new generated values are given to PSO for optimization

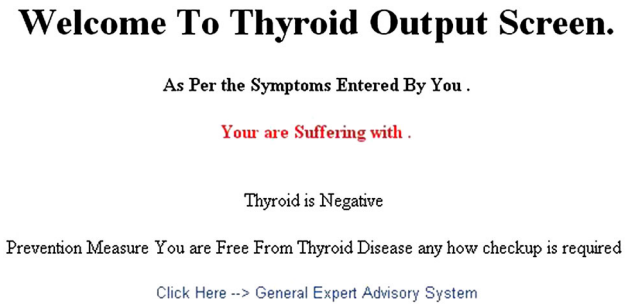

Fig. 6 Screen shot for PSO by using generated values of RDS displaying the optimistic TDD with prevention method

vention or curing method is done through the knowledge base.

If symptoms
S1 = 1, S2 = 1, S3 = 1, S4 = 1, S5 = 1, S6 = 1, S7 = 1, S8 = 0, S9 = 0, S10 = 1, S11 = 0, S12 = 1
Diagnose: T3 toxic
Prevention or curing method: Thyroxin levels are more and requested to compel down by removing the goitre if observed and to reduce the thyroxin levels by daily checkups radio iodine treatment.

CASE 2 :: RBS generates 100 % output.

If symptoms
S1 = 0, S2 = 1, S3 = 0, S4 = 0, S5 = 0, S6 = 0, S7 = 1, S8 = 1, S9 = 1, S10 = 1, S11 = 1, S12 = 0
String matching system Relevant data not found.
Then the string is passed to RBS and in *rule based system* relevant data is found
Diagnose: Goitre
Prevention or curing method: Swellings to be reduced and suggested antic radioactive ionized treatment.

CASE 3 :: SMS and RBS generates output as 0 %.
String matching system Relevant data not found,
Rule base system Relevant data not found.

If symptoms
S1 = 1, S2 = 1, S3 = 1, S4 = 1, S5 = 0, S6 = 0, S7 = 0, S8 = 1, S9 = 1, S10 = 1, S11 = 0, S12 = 0

Level 1: As per the considered data sets: (generated near by values through rough data sets theory → produces)

0.666667	0.666667	0.666667	0.666667
0.75	0.75	0.75	0.75
0.883	0.58	0.883	

Level 2: The values obtained through RDS are promoted for PSO → produces 98 % optimistic disease
Diagnose: Thyroxin levels are more
Prevention methodology: Prevention measure better reduce by anti thyroid medication or body metabolism changes (Figs. 3, 4, 5, 6).

5.1 Comparison of results for built systems SMS, ABC and PSO

See Table 5.

5.2 Comparison of results for built systems SMS, RBS, RDS, and PSO

See Table 6 and Figs. 7, 8 and 9.

6 Conclusion

Thyroid is one of the most common diseases. Since, thyroid datasets are uncertain data, missing attribute values and continuous features, the rough set theory treat these problems in the thyroid dataset. Moreover, The MSR, modified similarity analysis relation, is used to classify rules contains missing attribute value or gap, with respect to the number of the whole defined attributes for each rule. We have pre-

Table 5 When the data received from the user interface exactly matches the SMS system at outset then it produces the optimistic value exactly 1.0 which is 100 %, when SMS is success then the same result is obtained from individual algorithms ABC and PSO and if the SMS does not produces the relevant data, the input string is given to ABC algorithm to optimize, which produced a result rate from 0.65 to 0.75 which is 65 to 75 %, and similarly if ABC also found with Optimistic value less than 65 % or NIL, the input string is given to PSO which produced result 0.75 to 0.98 i.e. 75 to 98 %

Entered data	SMS	SMS-value	ABC	ABC-value	PSO	PSO-value
100000000000	DATA	1.0	DATA	1.0	DATA	1.0
110011001100	NO DATA	0.0	DATA	0.75	DATA	0.98
101010101010	NO DATA	0.0	DATA	1.0	DATA	1.0
111000111000	NO DATA	0.0	DATA	1.0	DATA	1.0

This individual algorithms workflow stops when the data produced consists of many missing values

Table 6 The comparative result between SMS and RBS, where SMS and RBS fails for sure due to missing value, but RDS theory generated the missing values and created a new knowledge base, which exactly when used with PSO algorithm, had produced 100 % accurate result

Entered data	SMS	SMS-value	RBS	RBS-value	RDS	RDS-value	PSO	PSO-value
110011001100	NO DATA	0	DATA	0	DATA	0.7	DATA	1.0
101010101010	NO DATA	0	DATA	0	DATA	0.7	DATA	1.0
111000111000	NO DATA	0	DATA	0	DATA	0.7	DATA	1.0
101111011110	NO DATA	0	DATA	0	DATA	0.7	DATA	1.0
100000111011	NO DATA	0	DATA	0	DATA	0.8	DATA	1.0
111110111111	NO DATA	0	DATA	0	DATA	0.8	DATA	1.0

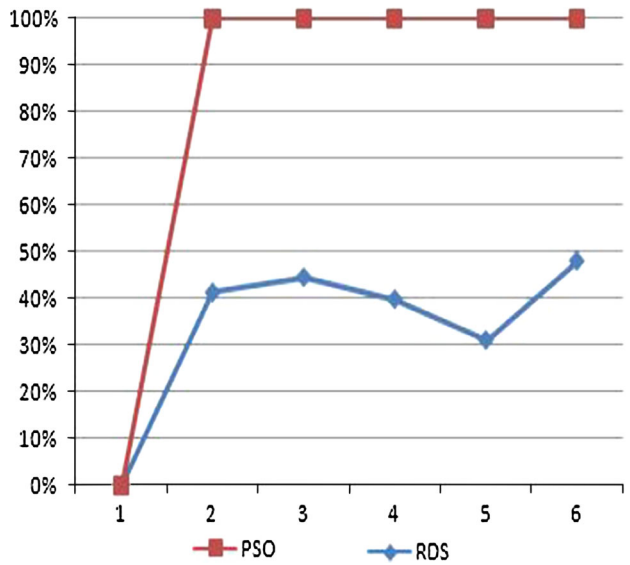

Fig. 7 The proposed comparison chart between RDS generated results and PSO results

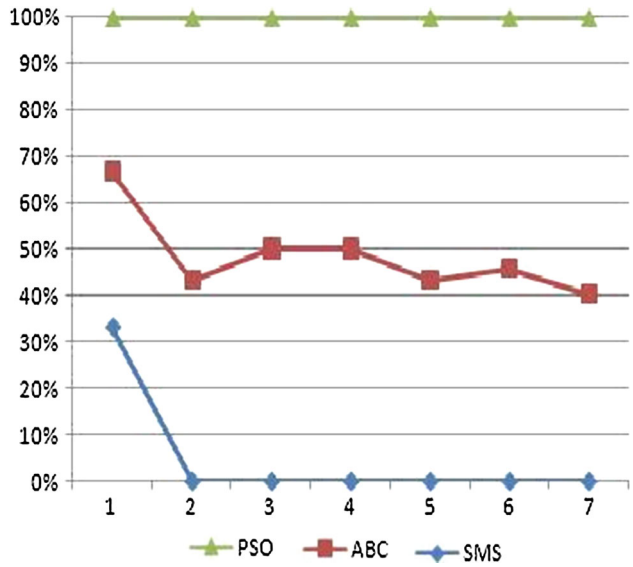

Fig. 8 The proposed comparison chart between SMS, ABC and PSO results

sented a reliable learning method and analytical studies for diagnosing thyroid disease that can be used by doctors in other medical diagnose algorithms. Indeed statistical results show that this evolutionary classification algorithm is the best in reducing size, time, attributes and increasing accuracy.

Although RST with MSR achieved good results that of the ML algorithms, it still suffer from unsatisfied accuracy measure. Hence a hybrid model of RST and ML algorithms should be introduced. This method uses the class information entropy of candidate or patients to select the bin boundaries. Moreover, the missing attribute values are treated based on computing the information gain by dropping an attribute, and then a similarity relation is measured. Hence, the Hybrid architecture system developed including rough data sets theory and machine learning algorithms in this work achieved an accuracy level for predicting the disease levels reached at most 99 to 100 % optimistic levels for the given data sets.

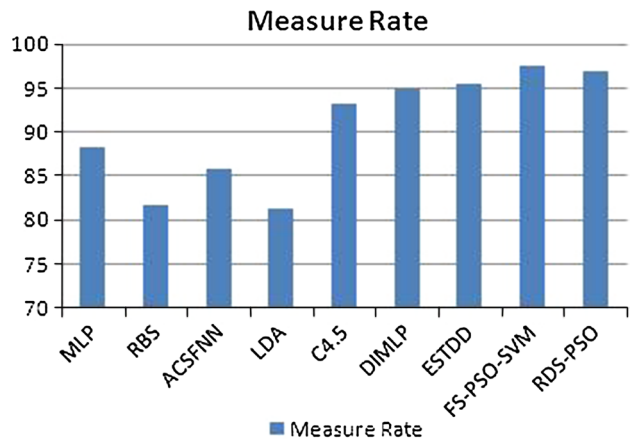

Fig. 9 This screen shot compares the thyroid disease diagnose obtained through several systems developed using data mining, image processing, artificial intelligence, expert systems and so on, and the systems developed having their working results up to 95 % considering a few attributes in identifying the major disease. This is the comparison chart produced based on my literature survey on several resolved algorithms like fisher core, learning patterns, design patterns, image patterns and so on

References

Akbas A, Turhal U, Babur S, Avci C (2013) Performance improvement with combining multiple approaches to diagnosis of thyroid cancer. Sci Res. doi:10.4236/eng.2013.510B055

Azar AT, Hassanien AE, Kim T (2012) Expert system based on neural-fuzzy rules for thyroid diseases diagnosis. Springer, Berlin

Crochemore, Czumaj et al (1994) Speeding up two string-matching algorithms. Algorithmica 12(4–5):247–267

El-Sayed El-Metwally S, Radwan E, Hamza T (2010) Multiple DNA sequence alignment using a hybrid model of GA and rough sets. Egypt Comput Sci J 34(3):46–60

Gharehchopogh FS, Molany M, Mokri FD (2013) Using artificial neural network in diagnosis of thyroid disease: a case study. Int J Comput Sci Appl 3(4):49–61

Guvenir HA, Demiroz G, Ilter N (1998) Learning differential diagnosis of erythemato-squamous disease using voting feature intervals. Artif Intell Med 13:147–165

Jacqulin Margre J, Lakshmipathi B, Aswani Kumar S (2012) Diagnosis of thyroid disorders using decission tree splitting trees. Int J Comput Appl 44(8):43–46

Karaboga D, Akay B (2009) A comparative study of artificial bee colony optimization. doi:10.1016/j.amc.2009.03.090

Karaboga D, Basturk B (2007) ABC optimization for solving constrained optimization problems. IFSA 2007. LNAI, vol 4529. Springer, Berlin

Keles A (2008) ESTDD: Expert system for thyroid diseases diagnosis. Expert Syst Appl 34(1):242–246

Li M, Zhou Z-H (2006) Improve computer-aided diagnosis with machine learning techniques using undiagnosed samples. IEEE 5:1–12

Pawlak Z (1991) Rough sets—theoretical aspects of reasoning about data. Kluwer Academic, Dordrecht

Peters G, Weber R, Nowatzke R (2012) Dynamic rough clustering and its applications. Appl Soft Comput 12:3193–3207

Polkowski L, Skowron A (eds) (1998) Rough sets in knowledge discovery, vols 1 and 2. Physica, Heidelberg

Prasad V, Rao TS, Babu MSP (2013) Offline analysis and optimistic approach on livestock expert advisory system. CiiT Int J Artif Intell Syst Mach Learn 5(12):488–495

Prasad V, Srinivasa Rao T, Veera Reddy A, Chaitanya B (2014a) Health diagnosis expert advisory system on trained data sets for hyperthyroid. Int J Comput Appl 102(3):40–46

Prasad V, Siva Kumar R, Mamtha M (2014b) Plug in generator to produce variant outputs for unique data. Int J Res Eng Sci 2(4):1–7

Radawan E, Assiri AMA (2013) Thyroid diagnosis based technique on rough sets with modified similarity relation. Int J Adv Comput Sci Appl 4(10):120–126