Hypothetical Description for Intelligent Data Mining

Madhu.G[1], G.Suresh Reddy[2]
Department of Information Technology,
VNR Vignana Jyothi Inst of Engg & Technology,
Batchupally, Nizampet (S.O), Hyderabad-90, INDIA.

Dr.C.Kiranmai[3]
Department of Computer Science & Engineering,
VNR Vignana Jyothi Inst of Engg & Technology,
Batchupally, Nizampet (S.O), Hyderabad-90, INDIA.

Abstract—**Intelligent data mining is to use the intelligent search to discover information within data warehouses those queries and reports cannot effectively reveal and to find the patterns in the data and infer rules from them, and use these patterns and rules to guide for decision making and forecasting. Therefore the Intelligent data mining incorporates advantages of both knowledge acquisitions from data, and knowledge acquisition from experts. In this paper, we propose a new framework on medical diagnosis for intelligent data mining using computations technique based on rough sets**.

Keywords: dengue fever; data mining; decision tree; rough sets;

I. INTRODUCTION

Data mining is the process of selecting, exploring and modeling large amounts of data to uncover previously unknown patterns [1]. Intelligent data mining requires a tight corporation between domain experts, in this case medical quality managers, and data mining experts and consists of data-driven as well as interest-driven analyses. The work is supported by our data mining tool, the *Knowledge Discovery Assistant* [2]. Knowledge discovery aims to extract high-level knowledge or create a high-level description from real-world data sets [3]. Soft computing techniques, involving neural networks, genetic algorithms, fuzzy sets, and rough sets are mostly widely used in the data mining phase of the overall Knowledge Discovery (KD) process. Fuzzy sets provide a natural framework for the process to deal with uncertainty [4]. Neural networks [5] and rough sets [6] are widely used for classification and rule generation. Recently few tools are used in intelligent data mining is case-based reasoning, neural computing, intelligent agents, and other tools like decision trees, rule induction, data visualization. Rough sets help in granular computation and knowledge discovery process. Data mining tools such as Genetic Algorithm(GA) are presently used to recognize patterns, anticipate changes, and learn the buying habits and preferences of electronic commerce customers in Internet-based transactions [7][8]. Commonly intelligent techniques are used in dengue fever analysis are fuzzy theory [9], decision trees [10], and Bayesian classifier [11].

A rough set is an intelligent mathematical tool for extracting knowledge from uncertain and incomplete data based information. The theory of rough sets can be used to find dependence relationship among data, evaluate the importance of attributes, discover the patterns of data, learn common decision-making rules, reduce all redundant objects and attributes and seek the minimum subset of attributes so as to attain satisfying classification. Moreover, the rough set reduction algorithms enable to approximate the decision classes using possibly large and simplified patterns [12] [13] [14] [15] [16] [17] [18].

In the following part of this paper, section 2, we present a brief description for rough sets concepts and its approximations. Section 3, describes the situation about the problem based on intelligent data mining and its architecture, some rules. The next part, section 4, is dedicated to experiment results and analysis.

II. RELATED WORKS

Recently many researchers various soft computing methodologies have been applied to handle the different challenges posed by the data mining [19]. The back bone of rough set theory is the approximation space and lower and upper approximations of a set. The approximation space is a classification of the domain of interest into disjoint categories.

A. Rough sets

The lower approximation is a description of the domain objects which are known with certainty to belong to the subset of interest, whereas the upper approximation is a description of the objects which possibly belong to the subset. Any subset defined through its lower and upper approximations is called a ***rough set***. The main advantage of rough set theory is that it does not need any preliminary or additional information about data – like probability in statistics, grade of membership in fuzzy set and so on.

B. Dengue fever data sets

The data sets are used in our experiments consists of 100 samples taken from different diagnosis laborites in Hyderabad and Mumbai from INDIA. Each sample consists of few measurements with label that denotes its class.

VNR Vignana Jothi Institute of Engg & Technology

***Definition* 1**: Information system is a tuple (U, A), where U consists of objects and A consists of features. Every $a \in A$ corresponds to the function $a : U \rightarrow V_a where V_a$ is the value set of a. In the applications, we often distinguish between conditional features C and decision feature D, where $C \cap D = \phi$. In such cases, we define decision system (U, C, D).

Patient	Attributes			
	Temperature	**Headache**	**Vomiting**	**Illness**
#1	High	No	Yes	Yes
#2	High	Yes	No	Yes
#3	Very High	Yes	Yes	Yes
#4	Normal	No	Yes	No
#5	High	Yes	No	No
#6	Very High	No	Yes	Yes

Table 1: Information table for Dengue Fever

The above table 1 classified into to that the set regarding {patient2, patient3, patient5} is indiscernible in terms of headache attribute. The set concerning {patient1, patient3, patient4} is indiscernible in terms of vomiting attribute. Patient2 has a viral illness, whereas patient5 does not, however they are indiscernible with respect to the attributes headache, vomiting and temperature. Therefore, patient2 and patient5 are the elements of patients' set with unconcluded symptoms.

***Definition* 2**: In rough sets theory, the approximation of sets is introduced to deal with inconsistency. A rough set approximates traditional sets using a pair of sets named the lower and upper approximation of the set. Given a set $B \subseteq A$, the lower and upper approximations of set $Y \subseteq U$ are defined as follows.

$$\underline{B}Y = \bigcup_{x:[x]_{B \not\subset x}} [x]_B$$

………………… (1)

$$\overline{B}Y = \bigcup_{x:[x]_{B \cap x \neq \phi}} [x]_B \quad \text{………………... (2)}$$

The positive region of X is defined as:

$$POS_C(D) = \bigcup_{X : X \in U / \overline{Ind_D}} \underline{C}\ X$$

……………………….. (3)

$POS_C(D)$ is the set of all objects in U that can be uniquely classified by elementary sets in the partition U / Ind_D by means of C [17]. The negative region $NEG_C(D)$ is defined by:

$$NEG_C(D) = U - \bigcup_{X : X \in U / Indo} \overline{C}\ X \quad \text{………… (4)}$$

is the set of all objects can be definitely ruled out as member of X. The boundary region is the difference between upper and lower approximations of set X that consists of equivalence classes having one or more elements in common with X; it is given by the following formula:

$$BND_B(X) = \underline{B}X - \overline{B}X \quad \text{…………….. (5)}$$

III. Rough sets on Intelligent Data Mining

Intelligent data mining is to use the intelligent search to discover knowledge within databases and warehouses those queries and reports cannot effectively reveal and to find the patterns in the data. Rough Set can be used in different phases of the knowledge discovery process, as attribute selection, attribute extraction, data reduction, decision rule generation and pattern extraction [19].

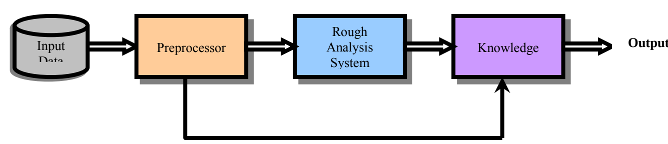

Fig.1. Architecture of Intelligent Data mining.

In this approach, input parameters will pass through the data into preprocessor and it pass a rough analysis system which will act as a data mining core for our system. Outputs of this system are appeared as a new database with some reductions in rows and columns. This means that redundancies in both attributes and entities of information system are discovered and omitted from the database. This block-set also recognizes condition attributes strongly affecting each decision one.

Rule 1:

If patient
blotched_red_skin=No and
muscular_pain_artculations = No and
temperature=Normal
Then dengue=No.

Rule-2

If patient
blotched_red_skin = Yes and
muscular_pain_articulations = No and
temperature = Very High
Then dengue = Yes.

Rule-3

If patient
blotched_red_skin = No and
muscular_pain_articulations = Yes and
temperature = High

Then dengue = Yes.

IV. Experiments and Analysis

In this section, we describe our experiment results, which are collected Dengue fever data from different medical diagnosis labs in Hyderabad, INDIA. Based on this data we created an information table, and information, it can generate the decision rules for the dengue diagnosis.

Patient Name	Conditional Attributes			Decision Attributes
	Blotched_red_skin	Muscular_pain	Temperature	Dengue Fever
P1	No	No	Normal	No
P2	No	No	High	No
P3	No	No	Very High	Yes
P4	No	Yes	High	Yes
P5	No	Yes	Very High	Yes
P6	Yes	Yes	High	Yes
P7	Yes	Yes	Very High	Yes
P8	No	No	High	No
P9	Yes	No	Very High	Yes
P10	Yes	No	High	No
P11	Yes	No	Very High	No
P12	No	Yes	Normal	No
P13	No	Yes	High	Yes
P14	No	Yes	Normal	No
P15	Yes	Yes	Normal	No
P16	Yes	No	Normal	No
P17	Yes	No	High	No
P18	Yes	Yes	Very high	Yes
P19	Yes	No	Normal	No
P20	No	Yes	Normal	No

Table 2: Dengue symptoms for the patients.

C. Imprecision coefficient $\alpha D(X)$**:** where αD is the quality of approximation of X, it's denoted by

$$\alpha D(X) = |D''(X)| / |D^*(X)| \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (10)$$

Where $|D''(X)|$ and $|D^*(x)|$ it represents the cardinality of approximation lower and upper, and the approximation are set $\neq\varnothing$. Therefore, $0 \le \alpha D \le 1$, if $\alpha D(X) = 1$, X it is a definable set regarding the attributes B, that is, X is crisp set. If $\alpha D(X) \le 1$, X is rough set regarding the attributes D. Then it apply for the Table 1, we get $\alpha D(X)$=3/5 for the patients with possibility of they are with Illness. Apply for the Table 2 using equation (10) for the patients with possibility of they are with dengue αD(X) =7/8; and also not with dengue αD(X) = 8/12.

D. Upper approximation $\alpha D(D^*(X))$**:** It is the percent of all the elements that are classified as belonging to X, it's denoted as $\alpha D(D^*(X)) = |D^*(X)| / |A|$ ……… (11)

From the table 1, we get $\alpha D(D^*(X))$= 5/6**,** for the patients that have the possibility of they be with illness.
Upper Approximation set (B*) of the patients that possibly have dengue are identified as D* = {P3, P4, P5, P6, P7, P9, P13, P18}
Upper Approximation set (B*) of the patients that possibly have not dengue are identified as D* = {P1, P2, P8, P10, P11, P12, P14, P15, P16, P17, P19, P20}
Using equation (11), for the patients that have the possibility of they be with dengue $\alpha D(D^*(X))$= 8/20, and for the patients that not have the possibility of they be with dengue $\alpha D(D^*(X))$=11/20**.**

E. Lower approximation $\alpha D(D''(X))$**:** It is the percentage of all the elements that possibility is classified as belonging to X, and is denoted as:

$$\alpha D(D''(X)) = |D''(X)| / |A| \quad \ldots\ldots. (12)$$

From table 1, $\alpha D(D''(X))$=3/6=1/2, for the patients that have illness.
Lower Approximation set (D") of the patients that are definitely have dengue are identified as B" = {P3, P4, P5, P6, P7, P13, P18}
Lower Approximation set (B") of patients that certain have not dengue are identified as
D" = {P1, P2, P8, P10, P12, P14, P15, P16, P17, P19, P20}
Using equation (12), for the patients that have dengue $\alpha D(D''(X))$ = 7/20, and for the patients that not have dengue $\alpha D(D''(X))$ = 8/20.

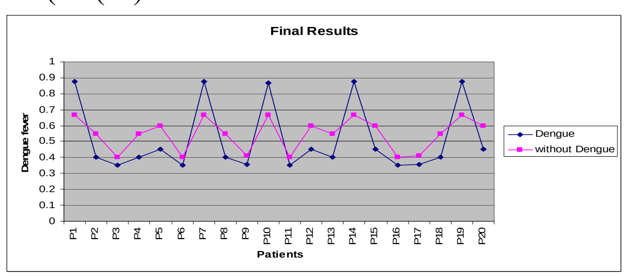

Fig.2. patients IgG & IgM values

Patient with dengue: $\alpha D(D''(X))$=7/20, that is, 35% of patients certainly with dengue. Patient that don't have dengue: $\alpha D(D''(X))$ = 11/20, that is, approximately 55% of patients certainly don't have dengue.10% of patients (P9 and P11) cannot be classified neither with dengue nor without dengue, since the characteristics of all attributes are the same, with only the decision attribute (dengue) not being identical and generates an inconclusive diagnosis for dengue.

V. CONCLUSIONS

In this paper, we presented an intelligent approach to data analysis with rough sets on data mining, this approach for the elimination of redundant data and the development of set of rules which can aid the doctor in the elaboration of the patient's diagnosis. Also process the incomplete data is based on the lower and upper approximations and theory was defined as a pair of the two crisp sets to the approximations. We derived information table which can be generated the necessary decision rules for the aid to the dengue diagnosis. The integration of rough sets with other intelligent tools such as fuzzy sets and neural network for classification and rule generation in soft computing paradigm is the aim of our future work.

REFERENCES

[1] S. Mitra and T. Acharya, *Data Mining: Multimedia*, *Soft Computing*, *and Bioinformatics*, John Wiley & Sons, Inc., NY, USA, 2003.

[2] Hogl, O.; Stoyan, H. et al.: *The Knowledge Discovery Assistant: Making Data Mining Available for Business Users*, in: Gunopulos, D.; Rastogi, R. (eds.): *Proceedings of the 2000 ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery (DMKD-2000)*, Dallas, Texas, May 2000, pp. 96-105.

[3] M. Kantardzic, *Data Mining: Concepts*, *Models*, *Methods and Algorithms*, John Wiley & Sons, Inc., New York, USA, 2002

[4] Y. Y. Liu and X. Q. Wu, "Evaluation for data fusion system based on uncertainty," *Journal of Data Acquisition & Processing*, Vol. 20, 2005, pp. 150155.

[5] A. Ataei, "Design of rubble mound breakwaters using artificial neural networks," Vol. M. Sc. Tarbiat Modares University, Tehran, Iran, 2002.

[6] P. Yang, "Data mining diagnosis system based on rough set theory for boilers in thermal power plants," *Frontiers of Mechanical Engineering in China*, Vol. 1, 2006, pp. 162-167.

[7] J. McCarthy, Phenomenal data mining, association for computing machinery, Communications of the ACM, 2000,43 (8): 75–80.

[8] T.K. Sung, N. Chang, G. Lee.Dynamics of modeling in data mining: Interpretive approach to bankruptcy prediction, Journal of Management Information Systems1999, 16 (1): 63–86.

[9] Parido, A., and P. Bonelli. A new approach to fuzzy classifier systems. In *Proceedings of the Fifth International Conference on Genetic Algorithms*. pp. 223–230. 1993.

[10] Hassanien, A.E. Classification and feature selection of breast cancer data based on decision tree algorithm. *International Journal of Studies in Informatics and Control Journal*, 12(1), 33–39.2003

[11] Cheeseman, P., and J. Stutz. Bayesian classification (AutoClass): theory and results. In U.M. Fayyad, G. Piatetsky–Shapiro, P. Smyth and R. Uthurusamy (Eds.), *Advances in Knowledge Discovery and Data Mining*. AAAI Press/MIT Press.1996.

[12] Grzymala–Busse, J., Z. Pawlak, R. Slowinski and W. Ziarko. Rough sets. *Communications of the ACM*, 38(11).1999.

[13] Kent, R.E. Rough concept analysis, rough sets, fuzzy sets knowledge discovery. In W.P. Ziarko (Ed.), *Proceedings of the International Workshop on Rough, Sets, Knowledge, Discovery*. Banff, Alta., Canada. Springer–Verlag. pp. 248–255.1994.

[14] Lin, T.Y., and N. Cercone (1997). *Rough Sets and Data Mining*. Kluwer Academic Publishers. Ning, S., H. Xiaohua, W. Ziarko and N. Cercone. A generalized rough sets model. In *Proceedings of the 3rd Pacific Rim International Conference on Artificial Intelligence*, Vol. 431. Beijing, China. Int. Acad. Publishers. pp. 437–443. 1994.

[15] Ning, S., H. Xiaohua, W. Ziarko and N. Cercone. A generalized rough sets model. In *Proceedings o f the 3rd Pacific Rim International Conference on Artificial Intelligence*, Vol. 431. Beijing, China. Int. Acad. Publishers. pp. 437–443.1994.

[16] Zhong, N., and A. Skowron. Rough sets in KDD: tutorial notes. *Bulletin of International Rough Set Society*, 4(1/2).2000.

[17] Polkowski, L., and A. Skowron. *Rough Sets in Knowledge Discovery*, Vol. 1/2. Studies in Fuzziness and Soft Computing series, Physica–Verlag. 1998.

[18] Polkowski, L., and A. Skowron. *Rough Sets and Current Trends in Computing*, LNAI 1424, Springer.Se-Ho, Ch., and P. Rockett (2002). The training of neural classifiers with condensed datasets. *SMCB*, 32(2),202–206. 1998.

[19] Sushmita Mitra, " Data Mining in soft computing framework: A Survey" IEEE Transactions on Neural Networks, Vol 13, No.1, January 2002

AUTHORS PROFILE

[1]**G.Madhu** completed his Master degree in Mathematics from J.N.T.University, Hyderabad in 2000 and his M.Tech degree in computer science & engineering from J.N.T.University, Hyderabad, INDIA, in 2008. Now pursuing PhD in Computer Science and Engineering from J.N.T.University, Hyderabad. He is presently working as Sr. Assistant Professor in Information Technology Department at VNR VJIET Hyderabad. His current research interest includes ANN, data mining, rough sets, and semantic web. He is a professional member of Indian Society for Rough Sets, and ISTE.

[2]**G.Suresh Reddy** Received the Bachelor's Degree in Computer Science and Engineering from Bangalore University Bangalore in 1997 and Master's Degree in Information Technology from Punjab University Punjab. Now pursuing PhD in Computer Science and Engineering from National Institute of Technology Warangal. He is presently working as Associate Professor and HOD in IT Department at VNR VJIET Hyderabad. His current research interests include Semantic Grid, Data Warehousing and Mining, Text Mining and Information Security.

[3]**Dr Mrs. C.Kiran Mai** did her graduate course in Electronics and Communication Engineering from Jawaharlal Nehru Technological University, Hyderabad and Masters in Software Systems from Birla Institute of Science and Technology, Pilani. From 1997 she is working in the Department of Computer Science, VNR VJ Institute of Engineering and Technology, affiliated to Jawaharlal Nehru Technological University. Currently she is the professor in the department. She handled the subjects - Microprocessors, Computer Networks, Visual Programming Techniques, Database Management Systems, Data Mining and Warehousing to undergraduate and graduate students of various disciplines. Before joining the teaching profession, she worked 7 years in industry where she was handling software projects in 'C', COBOL, Visual Basic and ORACLE. She is a life member of ISTE. Her research interests include Databases, Data mining, Networks and Human Computer Interaction. She is a member of the IEEE and the IEEE Computer Society.