IJCTA, 9(32), 2016, pp. 207-212
© International Science Press

ISSN: 0974-5572

An Intelligent Decision Support System for processing of medical data

K. Rajeswari*

ABSTRACT

The large collection of medical data available is not useful to the society unless the data is preprocessed and mined. Data preprocessing involves removal of noise, redundant data, understanding useful attributes and selecting relevant useful features. Data once preprocessed has to be mined for knowledge using association rule mining technique and soft computing techniques which include neural networks and genetic algorithm. This work uses the data from University of California, Irvine and real time data sets collected from Tanjore medical college and Madras medical college on heart disease and type 2 diabetes respectively. This work focuses on building an Intelligent Decision support system for processing the heart disease and type 2 diabetes data.

Keywords: data preprocessing, association rule mining, neural network, fuzzy logic, genetic algorithm, soft computing, heart disease, type 2 diabetes, decision support system.

1. INTRODUCTION

In the world full of data, intelligent processing can be applied whenever abundant data is available and is in need of analysis. Knowledge discovery process generally contains the following steps:

1. Management of Data.
2. Preprocessing of Data.
3. Data Mining tasks and Algorithms.
4. Post Processing.

According to Government of India statistics, the doctor - population ratio was a sparse 1:1722 in 2005. In this paper, the authors have proposed a solution for development of a decision support system with data preprocessing, association rule mining, and soft computing techniques.

The following Table 1.1 gives the ratio of Population Vs Number of cardiologists in different cities of India.

Table 1
Cardiologists Vs Population in India
(Source: Medical Council of India)

City	*Number of Cardiologists*	*Population*
Ahamadabad	67	5,835,000
Chennai	55	7,695,000
Mumbai	43	21,290,000

(*contd...*)

* Professor & Head, Department of Computer Engineering, Pimpri Chinchwad College of Engineering, Pune, India, *Emails: kannan.rajeswari@pccoepune.org, raji.pccoe@gmail.com*

(*Table 1 contd...*)

City	*Number of Cardiologists*	*Population*
Jaipur	33	4,245,000
Bangalore	29	7,365,000
Surat	22	4,265,000
Delhi	20	22,630,000
Calicut	18	439,922
Nagpur	18	2,665,000
Tirunelveli	18	3,072,880
All Cardiologists in India	538	1,21,01,93,422

Section 2 discusses about different methods and literature available for intelligent processing of data, Section 3 introduces our proposed decision support model. Section 4 gives detailed results and discussions and Section 5 concludes the paper.

2. RELATED WORK

Wu, et al. [1] have proposed a combination of decision support system with computer based data mining techniques to reduce the medical errors, improve patient's safety, reduce the unwanted practices and increase patients' true positive and true negative outcome. In this thesis, two types of data are dealt with.

1. Primary data collected from hospitals, through interviews with patients, laboratory data, physician's observations and interpretations.
2. Secondary data from benchmark UCI database[2].

Srinivas et al. [3] have discussed about disease classification or prediction results using Tree augumented Naïve Bayes network. Senthil Kumar[4] has used Fuzzy logic and Neural Network for diagnosis of heart disease.

Mohammed et.al. [5] have used association rule mining for horizontally partitioned databases for fully homomorphic encryption in privacy preservation. Jesmin et. al.[6] have made a study on sick and healthy factors closely associated with heart disease using association rule mining technique, Apriori.

According to Fitzpatrick and Grefenstette, Genetic algorithms are a search process that mimics natural selection. It is an evolutionary approach used for optimization problems. Ahmad et al.[7] has implemented a hybrid combination of Artificial Neural networks (ANN) and Genetic algorithm. The GA finds the optimal parameters and ANN classifies the test data accurately. Diabetes, Heart and Cancer data sets were used.

Han and Kamber[8] in their book on Data Mining have quoted many algorithms like ID#, C4.5 Probabilistic models, Neural based Back propagation and similar algorithms, along with their respective performance and applications. Das et al.[9] have introduced a technique to diagnose heart disease using SAS base software version 9.1.3. Experiments conducted on the data taken from Cleveland heart disease database have obtained a classification accuracy of 89.01%. The sensitivity and specificity values in heart disease diagnosis are 80.95% and 95.91% respectively.

Paper [10] summarizes the different datamining techniques used for medical datamining. Xu and Li (2010) confirmed that Intelligent Information Processing is a field of research for the past 10 to 15 years.

3. PROPOSED SYSTEM

This proposed work is formulated with the intention to seek out and discover interesting patterns of Ischemic heart disease and Type 2 Diabetes data. Figure 3.1 and Figure 3.2 depicts the complete framework of proposed methodology which has five major components as given below:

1. Data preparation
2. Patient data preprocessing
3. Frequent pattern generation
4. Data mining models for patient dataset
5. Knowledge representation

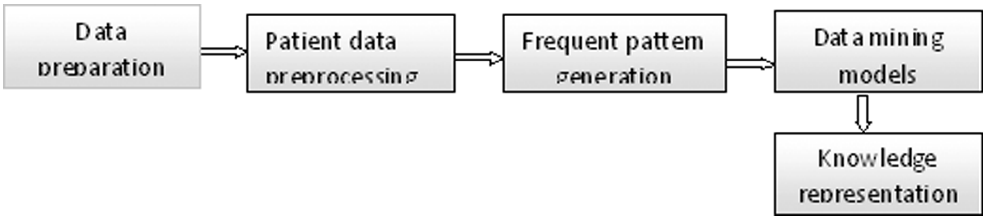

Figure 1: Proposed model

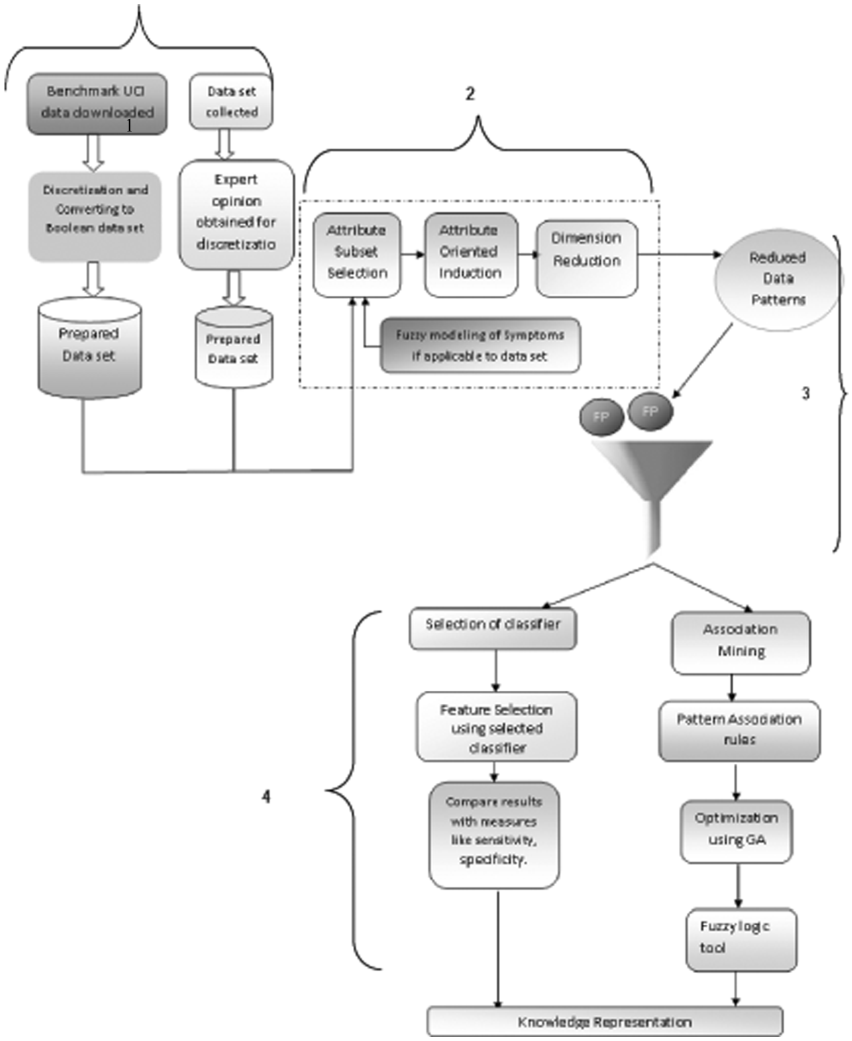
Benchmark UCI data downloaded
1
Data set collected
Discretization and Converting to Boolean data set
Expert opinion obtained for discretizatio
Prepared Data set
Prepared Data set
2
Attribute Subset Selection
Attribute Oriented Induction
Dimension Reduction
Fuzzy modeling of Symptoms if applicable to data set
Reduced Data Patterns
FP
FP
3
Selection of classifier
Feature Selection using selected classifier
Compare results with measures like sensitivity, specificity.
Association Mining
Pattern Association rules
Optimization using GA
Fuzzy logic tool
4
Knowledge Representation

Figure 2: Detailed Proposed model

3.1. Data Preparation

Data set from standard University of California, Irvine (UCI) is used. Three heart disease data sets namely Cleveland, Switzerland and Hungarian are collected from Cleveland Clinic Foundation, University hospital-Zurich (Switzerland) and Hungarian Institute of Cardiology (Budapset) respectively[2]. Real-time data sets are collected from Madras Medical College and Tanjore Medical College (refer Appendix).

3.2. Patient data preprocessing

As the medical data recorded by doctors is too complex, preprocessing is mandatory to transform the basic data in the form more appropriate for mining. Data reduction is used to acquire a diminishedrepresentation of the dataset similar in volume but still generates similar analytical results.At first, process is carried out to retrieve the task relevant attribute subsets that are sufficient for mining. Generalization is then done to make the attributes to split in a general manner, disregarding the individual differences. This generalized table will still occupy more space, memory and performance overhead. Hence dimension reduction is done on this generalized data set to remove unimportant attributes in order to make dataset amenable for mining.

3.3. Frequent pattern generation

This is a preliminary to identify frequently occurring symptoms from the pre-processed patient data set. As frequent occurrences may not possess any useful information, only patterns which satisfy a minimum frequency threshold are analyzed and extracted as frequent patterns.

3.4. Data mining models for patient data set

3.4.1. Models include

1. Patient classification as No risk, Medium risk and High risk
2. Association rules are generated by testing the association between candidate item-sets of frequent patterns, those satisfy minimum confidence threshold in order to measure their interestingness
3. Fuzzy modelling of symptoms whenever required like terms high fever, low pain, sever headache are quantified with qualitative knowledge available with experts.

3.5. Knowledge representation

The inferred knowledge from various mining models are consolidated and represented in textual as well as graphical form which will be more understandable to the users. Knowledge representation uses semantic phrases like, if-then-else rule structures, bar charts, line graphs and plain text notes. This knowledge will be kept in a knowledge base for future use and analysis.

4. RESULTS AND DISCUSSION

The use of a combination of techniques in this paper has been proved to be more effective than usage of classification or association data mining techniques alone. The 5-step frame work of (i) preprocessing, (ii) novel association rule mining, (iii) use Genetic algorithm to select high impact factors, (iv) fuzzy logic, (v) feature selection with learning algorithms and demonstrated the effectiveness when dealing with heart disease data and type 2 diabetes data classification.

The outcome of this research is presented below:

- The volume and space complexity of initial data set was massive and it was multifarious. Hence effectual preprocessing was done to reduce the space complexity. The space complexity was reduced to an extent of 71% after generalization and dimension reduction yield around 83% further.

- Association rule mining guarantees identification of occurrence of symptoms together with disease, its patterns are analyzed for different minimum support values.
- Fuzzy modelling improves the accuracy to 90% from 84.37%.

5. CONCLUSION

Thus a Decision support system was built with different soft computing techniques. Efficient data pre processing techniques has improved the quality of data and implementation of techniques on the data. Association rule mining has helped to identify the relationship between symptoms and disease. Soft computing techniques improve the knowledge obtained with selected features and optimized parameters. The future work is to use the model developed in this work to be extended further for different diseases like cancer, asthma, tuberculosis etc..In future, Decision support system in health care with Internet of Things(IoT) to help the needy in emergency can be established.

REFERENCES

[1] Wu R, Peters W, Morgan MW. The Next Generation Clinical Decision Support: Linking Evidence to Best Practice. Journal Health care Information Management 2002; 16(4):50-55

[2] Blake CL, Mertz CJ. UCI Machine Learning Databases. . http://mlearn.ics.uci.edu/databases/heartdisease/ 2004

[3] Srinivas K, Kavihta Rani B, and Govrdhan A. Applications of Data Mining Techniques in Healthcare and Prediction of Heart Attacks. International Journal on Computer Science and Engineering. 2010; 2(2): 250-255.

[4] Senthilkumar AV. Diagnosis of heart disease using fuzzy resolution mechanism. International Journal of Science and Applied Information Technology. 2012.

[5] Mohammed Golam Kaosar, Russell Paulet, Xun Yi. Fully homomorphic encryption based two-party association rule mining. Data & Knowledge Engineering. 2012; 76–78: 1–15.

[6] Jesmin Nahar, Tasadduq Imam, Kevin S. Tickle, Yi-Ping Phoebe Chen. Association rule mining to detect factors which contribute to heart disease in males and females. Expert Systems with Applications. 2013; 40: 1086-1093.

[7] Ahmad Mat Isa NA, Hussain Z, Osman MK. Intelligent medical disease diagnosis using improved genetic algorithm – multilayer perceptron network. Journal of Medical Systems. 2013; 37(2): 9934

[8] Han J, Kamber M. Datamining: Concepts and techniques. San Francisco, CA: Morgan Kaufmann Publishers. 2001

[9] Mohammed Abdul Khaleel, Sateesh Kumar Pradham, G N Dash, “A Survey of Data Mining Techniques on Medical Data for finding locally frequent diseases”, International Journal of Advanced Research in Computer Science and Software Engineering. 2013;3(8):149-153.

[10] Nahar A, Daasch R, and Subramaniam K. Burn-in reduction using principal component analysis. Proceedings of ITC, IEEE. 2005; 8; 10-155

APPENDIX: DATA SET

As a case study, data set from Madras Medical College was collected for classification of Ischemic Heart Disease. The dataset has 712 tuples and 16 classifying attributes with 1 class label. The data was collected in three stages,

Stage 1 with Physical Examination parameters.

Stage 2 with Co Morbid features and

Stage 3 with attributes about personal habits and hereditary details.

Table A.2.1.
Stage 1 of data for IHD

			Stage 1			
1. Age	2. Sex	3. Menopause	4. Height	5. Weight	6. Body Mass Index	7. Waist Measure

Table A.2.2.
Stage 2 - Co Morbid Factors of IHD

Stage 2 Co Morbid Factors				
8.SBP	9.DBP	10. Diabetes	11. Cholesterol	12.Thyroid

Table A.2.3.
Stage 3 of data for IHD

Stage 3				
12. Personal habits Y[1]/N[0]	13. Family history Y[1]/N[0]	14. Genetic factors Y[1]/N[0]	15. TypeA personality Y[1]/N[0]	16. Sleeping disturbance Y[1]/N[0]

The class label, a risk level of heart disease is given as in Table A.2.4 and Table A.2.5. It shows the diagnosis with value 0(no heart disease): <50% diameter narrowing and value 1(has heart disease): >50% diameter narrowing.

Table A.2.4.
Absolute risk for IHD/CAD

Absolute Risk for CAD			
No Risk	Low Risk	Medium Risk	High Risk

Table A.2.5.
Absolute 2 level risk for IHD/CAD

Absolute Risk for CAD	
No Risk	High Risk

TYPE 2 DIABETES DATASET

As a real time data, data set from Diabetes Research Center Tanjore was collected. The Diabetes dataset was collected and fuzziness in input attributes was modeled. Input attributes are collected in 4 stages.

Stage 1 has twelve preliminary symptoms as follows:

Polyurea, Polydipsia, Polyphagia, Nocturia, Tiredness/ Weakness, Loosing_weight, Giddiness, Ulcer, Sleeplessness, Itching, Shoulder pain.

Stage 2 has the following six input parameters:

Gender, Age, Height, Weight, Body Mass Index, No_of_ Children.

Stage 3 includes twelve input attributes namely:

Smoke_Type, Yrs_of_Smoking, Drink_Alcohol, Yrs_of_ Drinking, Freq_Drink, Food_Qty, Veg_NonVeg, NonVeg_Yrs, Freq_Nonveg, Grams, Heriditory, Other_relatives.

Stage 4 has the following 13 input attributes:

Fasting_Sugar, Postparandial_Sugar, Random_Sugar, Urea, Creatinine, Total Cholesterol, Low density Lipoprotein (LDL), High density Lipoprotein (HDL), Very Low Density LipoProtein(VLDL), Triglycerides, Uric acid, Glycated haemoglobin (Hb1A1C).