RESEARCH ARTICLE OPEN ACCESS

Role of Data Mining Techniques in Human Disease Diagnosis

Sunny Sharma
(Department of Computer Science, Guru Nanak Dev University, Amritsar, Punjab)

Abstract:

Medical informatics growth can be observed now days. Advancement in different medical fields discovers the various critical diseases and provides the guidelines for their cure. This has been possible only because of well heeled medical databases as well as automation of data analysis process. Towards this analysis process lots of learning and intelligence is required, the data mining techniques provides the basis for that and various data mining techniques are available like Decision tree Induction, Rule Based Classification or mining, Support vector machine, Stochastic classification, Logistic regression, Naïve bayes, Artificial Neural Network & Fuzzy Logic, Genetic Algorithms. This paper provides the basic of data mining with their effective techniques availability in medical sciences & reveals the efforts done on medical databases using data mining techniques for human disease diagnosis.

Keywords — **Decision tree Induction, Rule Based Classification or mining, Support vector machine, Stochastic classification, Logistic regression, Naïve bayes, Artificial Neural Network & Fuzzy Logic, Genetic Algorithms**

I. INTRODUCTION

Data mining has acknowledged an enormous deal of attention in field of Agriculture, Mathematics, Computer Science, Finance, Chemistry, Economics and especially in areas of Medical Sciences and Bio-informatics. Data mining is the process of analysing data from different perspectives and summarizing it into useful information that can be used to increase revenue, cuts costs, or both [14],[15]. DATA mining is the process of finding correlations, RELATION BETWEEN DATA or patterns. The main aim of this process is find the patterns thAT WERE PREVIOUSLY UNKNOWN [13], [12]. ONCE these RELATIONS are found these can be used in decision making. Multistep Approach involved in data mining which can be expressed as Fig. 1.

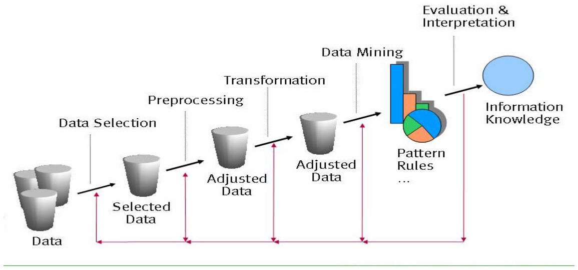

Fig. 1 Data mining Process

This Multistep Approach can be expressed as: Firstly data integration: At this initial stage data is being collected from all heterogeneous sources. Secondly data selection: At this stages select the data which is relevant one or in homogenous format. Then data cleaning: which is to remove the bugs or errors from data as data gathered is not in clean form. Then data transformation: which describe data after cleaning is not ready for mining as we need to transform data into forms appropriate for mining. These techniques are for smoothing, aggregation, normalization. Then data mining: now data is ready for applying techniques on data to locate or search the interesting patterns. Then pattern evaluation & knowledge presentation: This step involves visualization, transformation etc. finally deployment: which express decision /use of discovered knowledge [12].

II. DATA MINING TECHNIQUES

The data mining concept is originated from statistics, machine learning as well as from artificial intelligence. Lots of efforts have been done on data mining techniques. Broadly techniques in Data mining can be classified into two classes: Predictive

& Descriptive techniques. In predictive data mining the focus is done on discovering a relationship between independent as well as between dependent and independent variables. Predictive data mining can be used to forecast explicit values based on patterns in the data. Descriptive data mining describes a data set in a brief but comprehensive way and gives interesting characteristics of the data without having any predefined target. The various data mining techniques are shown in Fig. 2.

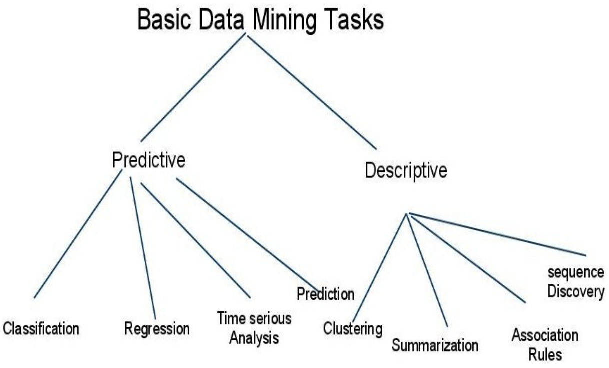

Fig.2: Techniques in Data mining

III. CLASSIFICATION

Classification is a data mining function that assigns items in a collection to target categories or classes. The goal of classification is to accurately predict the target class for each case in the data. Classifications are discrete and do not imply order. Continuous, floating-point values would indicate a numerical, rather than a categorical, target. A predictive model with a numerical target uses a regression algorithm, not a classification algorithm. Classification is a two step process: Learning or training Phase or Supervised Learning & Classification [20] [21].

Various Classification Methods are:

- Decision Tree
- Rule-based Methods
- Neural Networks
- Naïve Bayes
- Instance Based Learning
- Bayesian Belief Networks
- Support Vector Machines

A. Decision Tree Learning

This approach is machine learning inspired technique in which decision boundaries are explored while classification of data sets in the form of set of conditions on dependent variables. Domains of variables are noticed very carefully in this task. For considering many entrance doors of stimuli the normal conditions are to be tested firstly for deciding its contribution among various factors affecting it [13]. The Fig. 3 depict the decision tree approach.

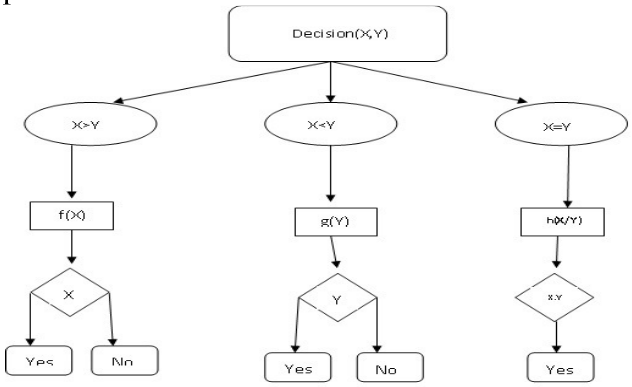

Fig.3: Decision Tree Learning

B. Artificial Neural Networks

Artificial neural network are biologically inspired machine learning technique in which a large networks of neurons exchange information between each other in order to produce knowledge. Neuron is the biological term used and it is analogous to nodes in the graph. Edges of these nodes carry weights which are tuned to particular path in order to favor learning process, which is same like natural learning process of human brain [14].

C. Support Vector Machines

Support vectors are the points in the vicinity of support hyper-plane. Support Hyper-plane is the best representation of data points which separates the large data sets into two classes. Support vectors are lying on the boundary of the plane favouring the classification of their domain contributing the plane.

D. Association Rules Learning

Association rules are the production rules which are applicable to perform relationship analysis between variables. There are strong association rules which are more likely to favour the strong relation between data items. Others are weak association rules which give least related items into consideration. An independent variable is appearing always in the right side of production rule while dependent variable is observed on the left which is also called antecedent and the former one is termed as consequent [22].

IV. Literature Survey

In 2010 A. K Banerjee et al. [5] did the stochastic classification on Protein database & describes the relation among physiochemical properties of proteins while keeping in view the hydrophobicity of AGC kinase related to super family. In 2010 A. A Deen et al. [4] express the power of Rule Based Classification by using the data set breast cancer, primary Tumor etc. Authors discuss the association rule based classification techniques & Naive bayes technique using various classification algorithms on Primary Tumor, Breast Cancer etc & attain the accuracy of 82%. In 2011 M. Singh et al. [8] describes the power of rule based mining & predict the classes of human protein function using sequence derived features. In 2011 J. Sony et al. [9] use the Genetic Algorithms, ANN, decision trees for the prediction of heart diseases & found decision trees outperform then others & it achieve the accuracy 99.2%. In 2012 A. Sudha et al. [11] predict the chance of stroke using central parameters & found ANN is better than Naïve bayes & decision trees in that process they attain accuracy of 91%. In 2012 M. Singh et al. [7] proposed the way to predict the protein function prediction from sequence derived features using See5 tool and design the decision tree through see5 decision rules & attain the accuracy of 64%. In 2013 S. RP. Singh et al. [10] predict staging of cervical cancer with genetic algorithms and describe the power of genetic algorithm. In 2014 A. Bhola et al.[6] successfully predict protein function through SDF using Support vector machine & attain the accuracy of 82.02%. In 2014 S. Saha et al. [3] predict protein function from protein interaction network using physicochemical properties. They combine the structural as well as sequential data information & express significance of sequence based properties over structure based approaches & achieves the accuracy 86%. In 2015 R. Singh et al. [2] use SVM (Support vector machine) a machine learning non linear classification & proposed the improved protein function classification using sequence extracted properties. In March 2015 M. A. Hussain et al [1] design a computer aided diagnosis (CAD) system using ANN & Fuzzy Logic to detect cancerous cell in CT scan images and did classification using ANN to detect the lung cancer. The comparative analysis of data mining techniques is expressed in Table1.

V. Results & Discussion

The important aspect of this paper is to highlight the role of data mining in medical science region, lots of data has already been mined in areas of protein, breast cancer, stroke, blood pressure, diabetic, & heart diseases etc. Variety of methodology & tools have used and they performed beyond the expectation & attain different level of accuracies. But research efforts are applied on little data sets rather than as a whole, different accuracies attain during research efforts are shown in Table II, and also their comparison shown in Fig. 4.

TABLE I: Comparative Analysis of Data Mining Techniques

Methodology & Tool	Year of Publication	Data Set Type/ Investigation Area	Author	Results Discussion
Decision tree Induction	July 2012	Protein Function Prediction	M. Singh et al. [7]	Authors proposed the way to predict the protein function prediction from sequence derived features using See5 tool and design the decision tree through see5 decision rules & attain the accuracy of 64%.
Rule Based Classification	2010	Breast cancer,	A. A Deen et	Authors discuss the association rule based

		Primary Tumor etc.	al. [4]	classification techniques using various algorithms on Primary Tumor, Breast Cancer etc & attain the accuracy of 82%.
	2011	Protein Function Prediction	M. Singh et al. [8]	Authors describe the immense use of rule based mining to predict the protein function.
Support vector machine	2015	Protein Function	R. Singh et al. [2]	Authors use SVM a machine learning non linear classification & proposed the Improved protein function classification using sequence extracted properties
	2014	Protein Function Prediction	A. Bhola et al.[6]	Author successfully predict protein function through SDF & attain the accuracy of 82.02%.
Stochastic classification	2010	Protein	A. K Banerjee et al. [5]	Authors describe the relation among physiochemical properties of proteins keeping in view hydrophobicity of AGC kinase super family.
Naïve bayes	2010	Breast cancer, Primary Tumor etc.	A. A Deen et al. [4]	Author discuss the association rule based classification techniques using various algorithms on Primary Tumor, Breast Cancer etc. & attain the accuracy of 82%.
	2012	Stroke	A. Sudha et al. [11]	Authors predict the chance of stroke using central parameters & found ANN is better than Naïve bayes & decision trees & they attain accuracy of 91%.
ANN & Fuzzy Logic	March 2015	Lung Cancer	M. A. Hussain et al [1]	Authors design a computer aided diagnosis (CAD) system to detect cancerous cell in CT Scan images and did classification using ANN
	June 2014	Protein Function Prediction	S. Saha et al. [3]	Authors predict protein function from protein interaction network using physicochemical properties. They combines the structural as well as sequential data information & express significance of sequence based properties over structure based approaches & achieve the accuracy 86%
Genetic Algorithms	2011	Heart	J. Sony et al. [9]	Authors use the Genetic Algorithm, ANN, decision trees for the prediction of heart diseases & found decision trees outperform & achieve the accuracy 99.2%
	2013	Cervical Cancer	S. RP. Singh et al. [10]	Authors predict staging of cervical cancer with genetic algorithms.

TABLE II: Research Effort Accuracies

Research Efforts	Level of Accuracy
M.Singh_ProteinPrediction	64
A.A.Deen_BreastCancer	82
A.Bhola_ProteinFunction	82.02
A.K.Banerjee_Protein	71.2
A.Sudha_Stroke	91
S.Saha_ProteinFunction	86
J.Sony_Heart	99.2

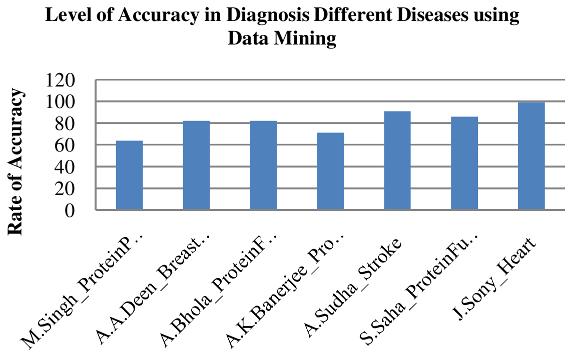

Fig.4: Research Effort Accuracies

VI. CONCLUSIONS

This paper provides a review over the efficacy of data mining in medical science region. These days protein, breast cancer, stroke, blood pressure, diabetic, & heart diseases are the core areas where lots of data has already been mined. Different methodology & tools have used and they performed beyond the expectation & achieve different level of accuracies. But research efforts are applied on little data sets rather than as a whole, different accuracies attain during research efforts. Still lots of areas are unrevealed like hypertension, physical disorder, dentistry, digestive disorder etc. and the authors did not consider other parameters for their real life research like family, person life style, etc. & still lots of efforts are needed for human health related problems.

REFERENCES

[1] M. A. Hussain, T. M. Ansari, P. S. Gawas and N. N. Chowdhury," Lung Cancer Detection Using Artificial Neural Network & Fuzzy Clustering," International Journal of Advanced Research in Computer and Communication Engineering, Vol. 4, Issue 3, March 2015.

[2] R. Singh , D.P kaur, "Improved protein function classification using support vector machine," International journal of computer science and information technologies. Vol. 6 (2015), 964-968.

[3] S. Saha, P. Chatterjee, "Protein function prediction from protein interaction network using physico-chemical properties of amino acids," International Journal of pharmacy and Biological Sciences, India, volume 4 issue 2 (2014),55-65.

[4] A. A Deen, M. Nofal, S. B. Ahmad, "Classification based on association rule based techniques: A general survey and empirical comparative evaluation," Ubiquitous computing and communication journal, Volume 5, No. 3, 2010.

[5] A. K. Banerjee and B. P. Manasa, "Assessing the relationship among physiochemical properties of proteins with respect to hydrophobicity: A case study on AGC kinase super family," Indian journal of Biochemistry & Biophysics, vol. 47, Dec. 2010, pp. 370-377.

[6] A.Bhola, S.K. Yadav and A.K.Tiwari, "Machine Learning based Approach for Protein function prediction using sequence derived properties," International journal of computer applications (2014), Vol. 105, 12.

[7] M. Singh, G. Singh, S. Sharma, "Human protein function prediction from sequence derived features using See5", International journal of scientific & engineering research, volume 3, issue 7, July 2012.

[8] M. Singh, G. Singh, Development of predictor for sequence derived features from amino acid sequence using associative rule mining, International journal of computer science and security (2011), vol. 5 issue 1.

[9] J. soni, U Ansari, "Predictive data mining for medical diagnosis: An overview of heart diseases prediction", International journal of computer applications, vol. 17, no. 8, 2011.

[10] S. RP. Singh, G. S. Randhawa, R. S. Virk, "Efficacy of genetic algorithms in staging of cervical cancer ", International journal of cancer research, vol. 47, issue 2, 2013.

[11] A. Sudha, P. Gayathri et al., "effective analysis & predictive model of stroke disease using classification methods", International journal of computer applications, vol. 43, no. 14, 2012, pp. 26-31.

[12] http://dataminingwarehousing.blogspot.in/2008/10/data-mining-steps-of-data-mining.html

[13] http://en.wikipedia.org/wiki/Data_mining

[14] http://www.zentut.com/data-mining/data-mining-techniques

[15] Jiawei Han and Micheline Kamber "Data Mining:Concepts and Techniques", 2nd Ed.

[16] http://www.cs.uiuc.edu/homes/hanj/bk2/toc.pdf.

[17] http://www.executionmih.com/data-mining/technology-architecture-application-frontend.php

[18] Mrs. Bharati M. Ramageri,"Data Mining Techniques and Applications," Journal http://www.ijcse.com/docs/IJCSE 10-01-04-51.pdf

[19] http://dataminingwarehousing.blogspot.in/2008/10/data-mining-steps-of-data-mining.html

[20] http://docs.oracle.com/cd/B28359_01/datamine.111/b28129/classify.htm

[21] http://ptucse.loremate.com/dw/node/13

[22] Chase Repp."Data mining" , http://www.uwplatt.edu/ChaseReppDataMining.doc