International Journal of Computer Science
Scholarly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
http://www.ijcsjournal.com Volume 5, Issue 2, No 02, 2017 ISSN: 2348-6600
Reference ID: IJCS-307 PAGE NO: 2053-2060

Classification Based Data Mining Technique for Discovering Disease Patterns: An Implementation Study

S.Baskaran

Head, Department of Computer Science

Tamil University

Thanjavur-613 010, India

sbaskarantj@gmail.com

R.Jeyaseelan

M.Phil Scholar, Department of Computer Science

Tamil University

Thanjavur-613 010, India

jayaseelanrayar@gmail.com

Abstract - Data Mining is getting popular in health care related research. It plays an important role for hidden information and disease patterns. The outcome helps for all those who have associated with heath care system. In this paper, we examine the application of classification based data mining techniques in healthcare data. Using medical profiles such as age, sex, pin code and disease type, it is predicted the likelihood of patients who get a large count of disease. The result of this quantitative minor research project helps to discover hidden patterns, associations between age and disease, and blood group and disease which often go unexploited.

Keywords— data mining, disease pattern, healthcare system, association mining,medical data mining

Introduction

Data mining can be defined as "the process of finding previously unknown patterns and trends in databases and using that information to build predictive models" [1]. In its simplest form, data mining automates the detection of relevant patterns in a database, using defined approaches and algorithms to look into current and historical data that can then be analyzed to predict future trends. The data mining results allow organizations to make proactive, knowledge-driven decisions and answer questions that were previously too time-consuming to resolve. Like other areas, the health sector has more need for data mining to find information from its data bases. The health care institutions apply data mining on their existing data to discover new, useful and potentially life-saving knowledge. Otherwise these types of hidden knowledge would have remained inert only in medical databases. By employing data mining and

All Rights Reserved ©2017 International Journal of Computer Science (IJCS Journal) and
Published by SK Research Group of Companies (SKRGC) - Scholarly Peer Reviewed Research Journals

International Journal of Computer Science
Scholarly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
http://www.ijcsjournal.com Volume 5, Issue 2, No 02, 2017 ISSN: 2348-6600
Reference ID: IJCS-307 PAGE NO: 2053-2060

visualization, medical experts could find patterns and anomalies better than just looking at a set of tabulated medical data. According to Jinn-Yi Yeh *et al.* (2011), data mining techniques can be broadly classified based upon what they can do, viz: (a) description and visualization; (b) association and clustering; and (c) classification and estimation; and thus can be a predictive modelling [2]. In this line, we did an experiment and discovered a set of knowledge. The discovered knowledge can be used by the medical analyzers for knowing the relationships in between disease and blood group, disease and age, disease and location etc. The outcome could help them for identifying the root causes of diseases, providing remedies and improving the health care management and thereby help society.

DATA MINING TOOLS

Number of and different types of off-shelf data mining tools are available commercially. Each tool has its own strengths and weaknesses. In general, Data mining tools can be classified into three categories: traditional data mining tools, dashboards, and text-mining tools. Traditional data mining programs help to establish data patterns and trends by using a number of complex algorithms and techniques. They are available in both Windows and UNIX platforms. Dashboards reflect data changes and updates onscreen. It is often in the form of a graph, chart or table. The inbuilt functionality makes dashboards easy to use. The third type of data mining tool sometimes is called a text-mining tool because of its ability to mine data from different kinds of text. These tools scan content and convert the selected data into a new but common format. This format is compatible with the tool's database. Capturing these inputs can provide a wealth of information that can be mined to discover trends, concepts, and attitudes. Besides these tools, other applications and programs are also available for data mining.

NEED OF ANALYSIS OF MASSIVE DATA IN HEALTH CARE SYSTEM

The health care organizations like hospitals, clinics are expected to have provision of quality services at affordable costs. This is a major challenge facing by them. Quality service involves diagnosing patients correctly and administering treatments that are effective. Poor clinical decisions cause disastrous consequences and so they are unacceptable. They can achieve the quality services with minimum cost by employing computer-based information and/or decision support systems. Computer based health care data is massive. It includes patient centric data, resource management data and transformed data. A number of relationships are hidden among such a large collection of data for example a relationship between patient data and their number of days of stay [3]. Data mining of this voluminous data would generate a knowledge-rich environment which can help to improve the quality of decisions and thereby quality of services in health care system. Data Mining is being used in the field of health to study genetic factors, environmental influences and physiological data allows practitioners to prevent, diagnose and treat diseases.

All Rights Reserved ©2017 International Journal of Computer Science (IJCS Journal) and
Published by SK Research Group of Companies (SKRGC) - Scholarly Peer Reviewed Research Journals

International Journal of Computer Science
Scholarly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
http://www.ijcsjournal.com Volume 5, Issue 2, No 02, 2017 ISSN: 2348-6600
Reference ID: IJCS-307 PAGE NO: 2053-2060

Such study helps to improve people's welfare. Application of Data Mining to clinical data promotes health and wellbeing. Data Mining of patient's data allow doctors to perform more accurate forecasts of diseases. This can keep older people active and independent for longer and help health and care systems to remain sustainable.

BRIEF DESCRIPTION OF OUR WORK

The following major obstacles for Data Mining in Healthcare are kept in mind while executing our work: [a] the raw medical data is voluminous and heterogeneous [b] content and complexity of medical data [c] missing, incorrect, inconsistent or non-standard data. For intelligent and effective disease analysis using data mining, we have done and presented the following.

(i) collected large transaction dataset
(ii) data mining goals are defined based on business intelligence and data exploration.
(iii) followed an efficient approach for the extraction of significant patterns from the disease data warehouses
(iv) followed an efficient approach for the efficient clustering of medical dataset

At the end, we found that the identified patterns and devised models could answer complex queries in predicting diseases.

APPROACH OF OUR WORK

Some of the approaches for incorporating subjective knowledge into the pattern discovery task are:

Visualization: This approach requires a user-friendly environment to keep the human user in the loop. It also allows the domain experts to interact with the data mining system by interpreting.

Template-based Approach: This approach allows the users to constrain the type of patterns extracted by the mining algorithm. Instead of reporting all the extracted rules, only rles that satisfy a user-specified template are returned to the users.

Subjective interestingness measure: A subjective measure can be defined based on domain information such as concept hierarchy or profit margin of items. The measure can then be used to filter patterns that are obvious and non-actionable.

Among them our approach for our work is Visualization.

MATERIAL AND METHODS

A. Data collection

The data base consists of a table with 5000 rows and eight columns (Sr.No, Name, Age, Sex, Mobile No, Pin Code, Blood Group, Disease): the 'Sr.No' column lists serial numbers that serve as a primary key for the table

B. *Tools and techniques*

Various types of data mining tools are currently available and each has its own merits and demerits. For our project, we have developed our own software for mining. It is used for investigating the patterns of diseases for each age group. The simple

All Rights Reserved ©2017 International Journal of Computer Science (IJCS Journal) and
Published by SK Research Group of Companies (SKRGC) - Scholarly Peer Reviewed Research Journals

International Journal of Computer Science
Scholarly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
http://www.ijcsjournal.com Volume 5, Issue 2, No 02, 2017 ISSN: 2348-6600
Reference ID: IJCS-307 PAGE NO: 2053-2060

graphical user interface of our software helps anyone to use the system for mining data for finding valuable hidden information, patterns, and insights.

1) Data selection: In this first stage of the mining process, the data are prepared and errors such as missing values, data inconsistencies, and wrong information are corrected.

2) Data preparation: The data preparation stage is crucial for data analysis. The software requires input in XLS format.

3) Data analysis: In this study, we used a associative technique for data mining.

4) Result database: At this stage, associated parameters have been chosen. The software processes the raw data and creates a result database in tabular form.

5) Knowledge evaluation and pattern prediction: This stage extracts new knowledge or patterns from the result database. An informative knowledge database is generated that facilitates pattern forecasting on the basis of prediction, probabilities, and visualization.

6) Deployment: The final stage of this process applies a previously selected model to new data to generate predictions.

EXPERIMENTAL ANALYSIS

The data that can be captured by a patient record are classified in three groups: a structured data, semi-structured data, and unstructured data [8]. Data mining consists of various methods. Different methods serve different purposes, each method has its advantages and disadvantages. Data mining tasks can be divided into descriptive and predictive [9]. A methodology known as association analysis is useful for discovering interesting relationships hidden in large data sets. The uncovered relationships can be represented in the form of association rules. A common strategy adopted by many association rule mining algorithms is to decompose the problem into two major subtasks:

1. Frequent Itemset Generation: The objective of this task is to find all the item-sets that satisfy the minsup threshold. These itemsets are called frequent itemsets.

2. Rule Generation: The objective of this task is to extract all the high-confidence rules from the frequent itemsets found in the previous step. These rules are called strong rules.

We obtained a patient database and conducted a data mining analysis on this data base using our software. Our method of Data Mining is descriptive. We enfold that the following variables can provide good indicators for identifying probable disease of patients: age, gender blood group and location. By mining five thousands of patient records, we quickly discovered several

All Rights Reserved ©2017 International Journal of Computer Science (IJCS Journal) and
Published by SK Research Group of Companies (SKRGC) - Scholarly Peer Reviewed Research Journals

International Journal of Computer Science
Scholarly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
http://www.ijcsjournal.com Volume 5, Issue 2, No 02, 2017 ISSN: 2348-6600
Reference ID: IJCS-307 PAGE NO: 2053-2060

hidden relationship among data. We furnished the discovered information in the following tables and corresponding visual graphs as the picture are easiest for people to understand and can provide plenty of information in a snapshot of the results. This facilitates health experts care to recognize patterns, to accept the basic trends in the data and formulate rational decisions.

The computational requirements for frequent itemset generation are generally more expensive than those of rule generation. Association Rules that can be derived from the above Tables of knowledge data are of the form:

1. (age, blood group) => Disease
2. (age, gender) => Disease
3. (age, blood group, location) => Disease
4. (age, gender, location) => Disease

Here, in our case:

Target Variable: Disease

Description of Predicator Variables:

Age: In years

Blood Group: A+, A1-,A1+, A1B-, AB-, AB+, B-, B+, O-, O+

Gender: Male or Female

Location: Town, Village

IF-THEN rules are specified as IF condition THEN conclusion e.g. IF age=old and blood group=B+ then Brain Tumor prone=yes.

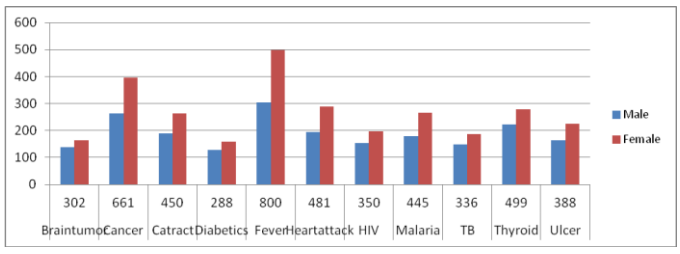
In this study, only a few relations and knowledge based conclusions could be made because of the lack of data in the database system in order to extract valuable associations and relations. Such kind of problems and drawbacks can be got over in further studies by including more fields or attributes in the data sets and by testing some other association algorithms.

Disease vs Gender

TABLE I: DISEASE vs GENDER

Disease	Total	Gender	
		Male	Female
Brain tumor	302	139	163
Catract	450	188	262
Diabetics	288	129	159
Fever	800	303	497
Heart attack	481	193	288
HIV	350	154	196
Malaria	445	180	265
TB	336	149	187
Thyroid	499	221	278
Ulcer	388	163	225
Total	5000	2083	2917

Fig 1: Bar Graph that illustrates Disease vs Gender



Disease vs Blood Group

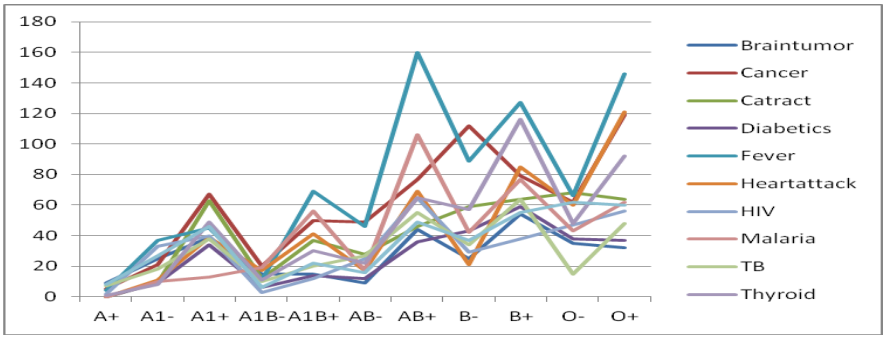
All Rights Reserved ©2017 International Journal of Computer Science (IJCS Journal) and
Published by SK Research Group of Companies (SKRGC) - Scholarly Peer Reviewed Research Journals
http://www.skrgcpublication.org/

International Journal of Computer Science
Scholarly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
http://www.ijcsjournal.com Volume 5, Issue 2, No 02, 2017 ISSN: 2348-6600
Reference ID: IJCS-307 PAGE NO: 2053-2060

TABLE II: DISEASE vs BLOOD GROUP

	A+	A1-	A1+	A1B-	A1B+	AB-	AB+	B-	B+	O-	O+	Total
Brain Tumor	9	25	39	15	15	9	44	25	54	35	32	302
Cancer	5	21	67	20	50	49	77	112	79	62	119	661
Catract	0	9	63	12	37	28	46	59	64	68	64	450
Diabetics	0	9	34	6	14	12	36	43	59	38	37	288
Fever	4	37	45	11	69	46	160	89	127	66	146	800
Heart Attack	0	11	39	17	41	17	69	21	85	60	121	481
HIV	2	33	40	3	12	25	65	29	38	47	56	350
Malaria	0	10	13	19	56	17	106	42	77	43	62	445
TB	7	18	38	10	20	27	55	34	64	15	48	336
Thyroid	1	8	49	11	30	22	65	57	116	48	92	499
Ulcer	8	27	46	6	22	16	49	37	55	62	60	388
Total	36	08	473	130	366	268	772	548	818	544	37	5000

Fig 2: Line Graph that illustrates Disease vs Blood Group



CONCLUSION

Association Analysis results should be interpreted with caution. The inference made by an association rule does not necessarily imply causality. Instead, it suggests a strong co-occurrence relationship between items in the antecedent and consequent of the rule. Causality, on the other hand, requires knowledge about the causal and effect attributes in the data and typically involves relationships occurring over time (e.g., Raining causes Mosquitoes that leads to Malaria).

All Rights Reserved ©2017 International Journal of Computer Science (IJCS Journal) and
Published by SK Research Group of Companies (SKRGC) - Scholarly Peer Reviewed Research Journals

International Journal of Computer Science
Scholarly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
http://www.ijcsjournal.com Volume 5, Issue 2, No 02, 2017 ISSN: 2348-6600
Reference ID: IJCS-307 PAGE NO: 2053-2060

Clinical repositories containing large amounts of biological, clinical, and administrative data are increasingly becoming available as health care systems integrate patient information for research and utilization objectives. The idea of health care data mining is to extract hidden knowledge in health care/medical field using data mining techniques. For computer professional, it is possible to identify patterns even if do not have fully understood the casual mechanisms behind those patterns. The identified patters, discovered relationships are helpful in studying the progression and the management of disease. A typical clinic data mining research including following ring: structured data narrative text, hypotheses, tabulate data statistics, analysis interpretation, new knowledge more questions, outcomes observations and structured data narrative text. Data mining programs lack the human intuition to recognize the difference between a relevant and an irrelevant data correlation, users need to review the results of mining exercises to ensure results provide needed information. The demonstration of our system to health care professional makes us to understand practically that "Mining of health records is not only good area of research, but also has its own efficiencies in the health care system for taking precautious measures". The results can help to identify and track patients and to design appropriate methods for lowering number of cases of diseases.

REFERENCES

1. E.W.T. Ngai, Yong Hu, Y.H. Wong, Yijun Chen and Xin Sun "The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature", Decision Support Systems, Vol. 50, pp. 559–569, 2011.
2. Jinn-Yi Yeh, Tai-Hsi Wu and Chuan-Wei Tsao "Using data mining techniques to predict hospitalization of hemodialysis patients", Decision Support Systems, Vol. 50, pp. 439–448, 2011.
3. H. C. Koh and G. Tan, "Data Mining Application in Healthcare", Journal of Healthcare Information Management, vol. 19, no. 2, **(2005)**.
4. Hastie, T., et al., 2009. The Elements of Statistical Learning, Data Mining, Inference and Prediction, 2nd edition. Springer, New York, USA.
5. Hosseinkhah, Fatemeh, et al. "Challenges in Data Mining on Medical Databases." (2009): 1393-1404.
6. Baylis, Philip. "Better health care with data mining." *SPSS White Paper, UK* (1999).
7. J. Nahar, T. Imam, K. S. Tickle and Y. P. Chen, "Association rule mining to detect factors which contribute to heart disease in males and females", Expert Systems with Applications, vol. 40, pp. 1086-1093, **(2013)**.

All Rights Reserved ©2017 International Journal of Computer Science (IJCS Journal) and
Published by SK Research Group of Companies (SKRGC) - Scholarly Peer Reviewed Research Journals

International Journal of Computer Science
Scholarly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
http://www.ijcsjournal.com Volume 5, Issue 2, No 02, 2017 ISSN: 2348-6600
Reference ID: IJCS-307 PAGE NO: 2053-2060

8. M. Kantardzic, Data mining: concepts, models, methods, and algorithms: Wiley-IEEE Press, 2003.
9. Bellazzi, R., & Zupan, B. (2008). Predictive data mining in clinical medicine: Current issues and guidelines.international journal of medical informatics 77, 81–97.

All Rights Reserved ©2017 International Journal of Computer Science (IJCS Journal) and
Published by SK Research Group of Companies (SKRGC) - Scholarly Peer Reviewed Research Journals