ISSN (Online) 2581-9429

IJARST

International Journal of Advanced Research in Science & Technology (IJARST)

Volume 1, Issue 1, June 2020

Smart Diseases Prediction System Using Data Mining

Ishank Darekar, Siddhi Desai, Prachi Govekar, Prof. Satish Ranbhise

Department of Information Technology

Shree L. R. Tiwari College of Engineering Thane, India

darekarishank@gmail.com and siddhidesai1998@gmail.com

siddhidesai1998@gmail.com and satishranbhise@slrtce.in

Abstract: Medicines generates a strong deal of data stored in the medical database. Deriving useful knowledge and making scientific decision for diagnosis and treatment of disease from the database increasingly becomes important. Data mining in medicine can help to deal with this situation. It can also improve the management level of clinical information and endorse the development of community and telemedicine. The main objective of this system is to use the medicinal data and the algorithms are run on that information and result will be displayed in the form of user understandable words and graph. When very large data sets are present, data mining algorithms (here considering only ID3) is used. ID3 outputs the result in the form of decision tree which can be easily understood by end user. Recommender systems aim to provide users with personalized products and service to deal with the increasing online information overload problem.

Keywords: Data mining, medical, disease, ID3

I. INTRODUCTION

Information technologies in health care have made provision for creation of patient records obtained from monitoring of patient visits. This information includes type of disease, patient information, lab results, etc. Health records are private information, yet the use of these sensitive documents may help in treating various diseases. Before data mining process begins, healthcare organizations must formulate a clear policy concerning privacy and security of patient records. The policy must be completely implemented in order to ensure patient privacy. The problem of prediction in medical domain can be divided into two phases: learning phase and decision making phase. In the learning phase, a large dataset is transformed into a simplified dataset. Number of features and objects in this new set are much smaller than the original set in several different ways. The rules generated in this phase are used to make accurate decisions based on dataset. Newly formed data set is used to make predictions when the new instances with unknown outcomes occur with the predictive algorithm. A huge amount of medical records are stored in databases and data warehouses. Such databases and applications differ from one another. With the evolution of machines, we have found that some time consuming and complex mathematical calculations can be done using calculators. Using current machines, specific information in a large data set can be found very fast and in an easy manner. We use different machines for storing information; reminding us of appointments, and so on. As the size of the data is increasing, computer storage also increases. Due to the vast amount of data that has been created, algorithms were invented which produce results once a query is supplied.

II. EXISTING SYSTEM

The Abbreviated Injury Scale (AIS) algorithm is the first published algorithm to produce all large item-sets in a transaction database. The AIS is an anatomical based coding system created by the Association for the Advancement of Automotive Medicine to classify and describe the severity of specific individual injuries. This algorithm has targeted to discover qualitative rules. This technique is limited to only one item in the consequent. This algorithm makes multiple passes over the entire database. In AIS, the frequent item sets are generated by scanning the databases several times. The support count of each individual item is accumulated during the first pass over the database. Based on the minimal

IJARST
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science & Technology (IJARST)
Volume 1, Issue 1, June 2020

support count, the items whose support count is less than its minimum value gets eliminated from the list of items. Candidate 2 item sets are generated by extending frequent 1 item set with other items in the transaction. During the second pass over the database, the support count of these candidate 2-itemsets Data security and privacy are the important issues when health related data is considered. Thus Health informatics deals with biomedical information and knowledge along with their storage, retrieval and optimal usage for problem solving and decision making. One of the unique characteristics of medical data mining is that, the result can be retrieved in the form of description of words, pictures or in graphical format (as like bar charts, pie charts, etc). In data mining, data set plays very important role. Data set can be taken for any particular disease or group of diseases.

III. PROBLEM STATEMENT

Clinical decisions are often made based on doctors instincts and experience rather than on the knowledge- rich data hidden in the database. Sometimes this practice leads to unwanted biases, errors and excessive medical costs which affects the quality of service provided to patients. Wu, et al put forth that integration of clinical decision support with computer-based patient records could reduce medical errors and unwanted practice variations, increase patient safety and improve patient medical result. This advice is encouraging as data modelling and analysis tools, e.g., data mining have the capability to generate a knowledge-rich environment which can help to significantly enhance the quality of clinical decisions.

IV. PROPOSED SYSTEM

Decision Tree is one of the most popular classification algorithms in currently used in Data Mining. Decision Tree includes various types of algorithms such as ID3, C4.5, C5, J48 and CART. In this system, ID3 algorithm is utilized because it best suited for heart disease dataset. The basic idea of Iterative Dichotomiser 3 or ID3 algorithm is to construct the decision tree by employing a top-down, greedy search through the given sets to test each attribute at every tree node. The main concepts used in ID3 algorithm are Information Gain and Entropy to select the attribute that is most useful for classifying the given sets. Entropy is defined as the measure of the amount of uncertainty in the (data) set S (i.e. entropy characterizes the data set).

Equation:
Entropy is calculated by

$$\text{Entropy}(S) = \Sigma\, n = 1\text{- } p(I) \log_2 p(I).$$

where, p(I) refers to the proportion of S belonging to class I
i.e., in our dataset we have two kinds of class {Yes, No}.
Information gain IG(A) is the measure of difference between entropy in S before split and after split on an attribute A. In other words, how much uncertainty in S is reduced after splitting set S on attribute A.

Information Gain is calculated by

$$\text{Gain}(S, A) = \text{Entropy}(S) - \Sigma((|S_a|/|S|) \text{ x Entropy}(S_a))$$

where, S is the total collection of records.
A is the attribute for which gain is calculated.
v is all the possible values of the attribute A.
S_a is the number of elements for each v.
Σ is the summation of $((|S_a|/|S|) \text{ x Entropy}(S_a))$ for all the items from the set of v.

The steps involved in ID3 algorithm are:

1. Calculate entropy of every attribute using dataset.
2. Split the sets into subsets using the attribute for which entropy is minimum or (equivalently, information gain is maximum).
3. Make a decision tree node containing the attribute.
4. Recurs on subsets using remaining attributes.

IJARST
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science & Technology (IJARST)
Volume 1, Issue 1, June 2020

Data mining basically consists of 4 types of techniques.
They are:

1. Classification
2. Association
3. Sequencing
4. Clustering

In this system we are making use of classification techniques Classification is a data mining (machine learning) technique used to predict group membership for data instances. Classification predicts categorical labels and prediction models continuous valued functions. Classification is the process of generalizing the known structure to apply to new data. Classification routines in data mining uses variety of algorithms and the particular algorithm used can affect the way records are classified. A usual approach for classifiers is to make use of decision trees to partition and segment the records. New records can be classified by traversing the tree from the root of the record through branches and nodes, till the leaf representing a class. The path record takes through a decision tree can then be represented as a rule.

Data mining can be advantageous as it can help

- Healthcare insurers detect fraud and abuse.
- Physicians identify effective treatments and best practices.
- Patients receives better and inexpensive healthcare services.
- One of the key advantages of using data mining is their speed in working with large data sets. Generation of quick results and faster analysis that can increase operational capability and reduce operating cost.
- Data Mining can extract predictive information from large database sources which is a very primary feature of Data Mining.

Disadvantages

- Heterogeneity of medical data Volume and complexity, Physician's interpretation with Poor mathematical categorization.
- Ethical, Legal and Social Issues
- Data Ownership Lawsuits
- Privacy issues
- Security issues
- Misuse of Information/inaccurate information

V. ANALYSIS AND PLANNING

The analysis phase is the most crucial phase of any project. The quality of the analysis can make or break a project. Artificial intelligence presents challenges due the complexity involved in getting the balance between too much and not enough. Planning such a task is matter of finding what we really need and we don't. We may design systems which can process an extremely wide variety of inputs, but we cannot actively ensure that the AI will respond to it in the way we want it to. It may present us an output which was intended to be for another input, or it may not be able to process it due to the load caused by the heavy processing on the interpreter. We have to plan for such a scenario too. Thus, the challenge is not just of input handling or building features, but of efficiency as well. We did this phase slowly so we be could be thorough with all our requirements and plans. A slow approach also allowed us to anticipate risks. Planning phase is probably the best time to plan for risks and avoid them altogether. Being careful in the planning phase allows us to put less effort into the risk mitigation, risk monitoring and risk management plan.

Feasibility Analysis

The primary phase in any system's developing life cycle is introductory investigation. The feasibility study is a major part of this phase. A measure amount of how profitable or practical the development of any system would be to

IJARST
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science & Technology (IJARST)
Volume 1, Issue 1, June 2020

the organization is the feasibility study. The feasibility of the development software or system can be studied in terms of following aspects:

1. Operational Feasibility.
2. Technical Feasibility.
3. Economic feasibility.
4. Motivational Feasibility.
5. Legal Feasibility.

Operational Feasibility:

The system will minimize the time consumed to maintain manual records and is not boring and unmanagable to maintain the records. Hence operational feasibility is assured.

Technical Feasibility:

- At least 1GHz Pentium Processor or Intel compatible processor.
- At least 1GB RAM.
- 14.4 kbps or higher modem.
- A video graphics card.
- A mouse or other pointing device.
- Atleast 1 GB hard disk space.
- Microsoft Internet Explorer 8.0 or higher.

Economic Feasibility:

Once the hardware and software requirements gets satisfied, there is no need for the user of our system to spend money for any additional overhead. The system will be economically feasible in the following aspects:

- The web site will reduce a lot of paper work. Hence the cost will be reduced.
- Our web site will reduce the time that is wasted in manual processes.
- The storage and managing issues of the registers will be solved.

Legal Feasibility:

The licensed copy of the required software is quite cheap and easily available. So from legal aspect the proposed system is legally feasible.

VI. CONCLUSION

The number of people getting sick and being admitted into clinics and hospitals are increasing proportionally. The growing number of patients indirectly increases the amount of data that are required to be stored. As the size of data increases, computer storage also increases. Due to the vast amount of data that has been created, humans invented algorithms that produce results once a query is supplied. The goals that have been achieved by the developed system are:

1. Simplified and reduced manual work.
2. Large volumes of data can be stored.
3. It provides Smooth workflow.

IJARST
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science & Technology (IJARST)
Volume 1, Issue 1, June 2020

REFERENCES

[1]. Bara, Adela, and Ion Lungu. "Improving decision support systems with data mining techniques." Advances in Data Mining Knowledge Discovery and Applications. InTech, 2012.

[2]. Burn-Thornton, Kath E., and Simon I. Thorpe. "Improving clinical decision support using data mining techniques." AeroSense'99. International Society for Optics and Photonics, 1999.

[3]. Musen, Mark A., Blackford Middleton, and Robert A. Greenes. "Clinical decision-support systems." Biomedical informatics. Springer London, 2014. 643-674.

[4]. Kawamoto, Kensaku, et al. "Improving clinical practice using clinical decision support systems: a systematic review of trials to identify features critical to success." Bmj 330.7494 (2005): 765.

[5]. Chang, C. C. and H. M. Lu, (2009), "A SOA-Based Medical Diagnosis Decision Support System using the Bayesian theorem and web service technology", Journal of the Chinese Institute of Engineers, 32(7), 923-930.

[6]. H.W. Ian, E.F., "Data mining: Practical machine learning tools and techniques," 2005: Morgan Kaufmann.

[7]. R. Detrano, A.J., W. Steinbrunn, M. Pfisterer, J.J. Schmid, S. Sandhu, K.H.Guppy, S. Lee, and V. Froelicher, "International application of a new probability algorithm for the diagnosis of coronary artery disease," American Journal of Cardiology,1989. 64: p. 304-310.

[8]. G. John, "Models if incremental concept formation," Journal of Atificial Intelligence, 1989: p. 11-61.

[9]. Senthil Kumar, B & Gunavathi R., Dr. (2016). A Survey on Data Mining Approaches to Diabetes Disease Diagnosis and Prognosis. IJARCCE. 5. 463-467. 10.17148/IJARCCE.2016.512105.

[10] Senthil Kumar, B & Sreejith.R. (2016)."A Survey on Identification of Diabetes Risk Using Machine Learning Approaches". IJIRCCE. 4. 33 -335 . 10.15680/IJIRCCE.2016. 0408001.