International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Enhanced and Improved Disease Prediction System in Healthcare Datasets Using Machine Learning Techniques

Shakeel Juman T. P.[1], Dr. K. Aravinthan[2]

Submitted: 16/10/2023 **Revised**: 06/12/2023 **Accepted**: 16/12/2023

Abstract: Healthcare uses Data Mining techniques (HDMT) can be for better application of knowledge and identifying successful prescription patterns for diseases. Usage of computer aided diagnosis for expert opinion learning has definite advantage. Integrated. Data Mining (IDM) with forecasting can provide a dependable and a high-quality desirable outcome. Prediction of diseases using data mining techniques is a motivating task for augmenting diagnostic accuracy. Hence the objective of this research is usage of HDMT/IDM methodology that can take less time and which can be more economical. The methodology can be useful to predict healthcare diseases. Hence to understand the usage of this research work is to identify the methodology to predict Healthcare diseases from patient's records and suggest a non-invasive data mining model.

Keywords: *HDMT, IDM, CART, Clustering, Diagnosis, Data mining, Decision Making, Healthcare.*

1. Introduction

Healthcare delivery system generates and stores enormous quantum of primary data. While technological advancements in the form of computer-based patient record software and personal computer hardware are making the collection of and access to health care data more manageable, few tools exist to evaluate and analyze this clinical data after it has been gathered and filed.

Analysis of primary data available can enhance the better management of disease progression. Better search modes need to be developed for the task. Past efforts in this area have been limited primarily to epidemiological studies on a Data Mining initiative and claims databases. Discovery in Databases or Knowledge Discovery in Databases (KDD), is the search for relationships and global patterns that exist in large databases but are 'hidden' among the vast amounts of data [1]. The typical data mining process involves transferring data originally collected in production systems into a data warehouse, cleaning or scrubbing the data to remove errors and check for consistency of formats, and then searching the data using statistical queries, neural networks, or other machine learning tools [2]. Though many applications of KDD have focused on discovering novel data patterns to solve business related problems, they have also been used extensively in the healthcare study and researches. Data mining has been used is used to discover subtle factors affecting the success and failure of therapeutic modalities which led to improvements in patient care [3].

MATLAB can work with matrices, deleting a row, a column, transposing a matrix, calculating the determinant etc. Similarly, Data Mining systems can be used for identifications and intervention strategies of diseases that were likely to cut costs and thereby reducing the economic burden. Thus, the eventual goal of knowledge recovery effort is to identify factors that can improve the quality and reduce costs in mining the healthcare information. This research work analysis the data mining techniques which are tested with MATLAB.

2. Problem Statement

Human population explosion has resulted in the manifestation of many novels and hither by undocumented diseases, where certain diseases may not have a permanent cure. Treating diseases are a major challenge to Health care providers as the Signs and symptoms of disease simulate other disorders. Management process in the healthcare facility is a major challenge to healthcare providers, thus making complex issue. A patient's feelings of distress, guilt, and anxiety occur due to their negative social experiences when they fail to understand the intensity of their illness [5]. Moreover, patients experience unique challenges personally, when diagnosed with an invisible illness that threatens their quality of life [6]. To improve a patients' emotional and physical health, awareness on healthcare diseases in patients and proper direction in diagnosis of disease and management are needed. Patients are vulnerable to psychological apprehension due to the decreased quality of living. Patients need treatment quickly and prompt intervention to reduce the morbidity to identify and take care of exhibited effectively. Thus, though disease can be treated effectively with the modern parts, the diagnosis early is a major problem due to its manifestations simulates other diseases. In processing data for identifying disease, machine learning techniques need previous history of patients adequately. Though current improvements in Medicare have enhanced patient's survival rates, fear of morbidity and mortality persists.

3. Proposed Work

- To get an early warning signal of an impending diseases

[1]*Ph.D Research Scholar, Department of Computer Science, Adaikalamatha College, affiliated to Bharathidasan University Vallam, Thanjavur, Tamilnadu*
shakeeljuman81@gmail.com
[2]*Assistant Professor, Department of Computer Science Adaikalamatha College, affiliated to Bharathidasan University Vallam, Thanjavur, Tamilnadu*
aravinthk83@gmail.com

- To reduce errors in diagnosis and management.
- To design and propose new algorithms and techniques to predict diseases early.
- To analyze existing algorithms and compare the proposed method's effectiveness.
- To reduce the time lag in prediction there by reducing the cost of the management.
- To extend the life span with improved quality of life in an afflicted patient.

4. Methodology

The data storage, processing and retrieval can enhance the better management for which complete comprehensive details of every individual is needed.

The following methodology is used

1. Dataset
2. Data Cleaning
3. Missing Value Prediction
4. Dataset Preparation
5. Decision Making

A. Dataset

The Dataset used in this work is the Chicago Lupus Database (CLD). This is a registry of individuals with lupus used for lupus research with a probable or definite lupus symptom. CLD attributes are Age, Gender, Test Sample, Disease Activity, Symptoms, Severity, Involved Organs, Tests conducted and follow-ups.

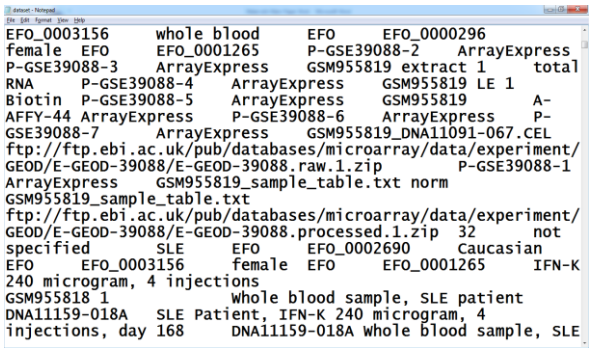


Fig 4.1. Sample dataset

B. Data Cleaning

Data cleaning is an important part of machine learning for its significance in model building. Data cleaning can make or break an analysis. Professional data analysts spend a lot of time in this step. A clean dataset can get desired results even with a simple algorithm, which is beneficial. Figure 4.2 depicts the flow of Data Cleaning.

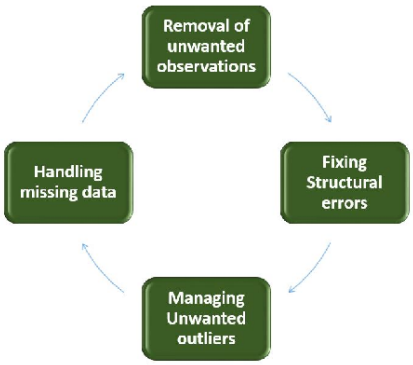


Fig 4.2. Data Cleaning Flow

Data cleaning involves different steps for different data. The step followed in this research work is Missing Value Prediction, Redundancy Avoidance, Filtering (Fill mean mode value) and Attribute Reduction.

C. Missing Value Prediction

Missing data can be identified in three ways as detailed below

- Missing Completely at Random (MCAR): Random missing values are the highest level of randomness where features are not dependent on any other features values.
- Missing At Random (MAR): Values missing in features dependent on other feature values.
- Missing Not at Random (MNAR): This implies the data gathering process has to be checked.

D. Dataset Preparation

A dataset is a comprehensive collection of patient's data. The data set corresponds to the contents of a single database table, or a single statistical data matrix, where every column of the table represents a particular variable, and each row corresponds to a given member of the data set. Machine Learning methods use a training data set where actual data is used to train the proposed model for performing various actions. The training data set applies concepts like neural networks for learning and expected results. It includes both input and expected output data. Training sets make up the majority nearly 70% of the total data. The testing model adapts to fit to parameters in a process called adjusting weights. The test data set is then used to evaluate how well a machine learning technique was programmed with the training data set. Testing sets represent remaining 30% of the data and ensures the input data grouped together is verified with correct outputs.

E. Decision Making

Making the right decision is often a challenge. A simple and quick approach for taking a decision is following past experiences in similar situations. The human brain decision is based on two factors namely logical and intuitive. Most decision are an automatic response as the logical part invents a reason for the decision. The intuitive system based on an entity from several plausible conclusions which need to be assessed. Tools like decision matrix can

help in unbiased decision making. It is an advanced approach for making decision and scores each possible option against certain criteria or feature. This approach results in creating decision matrix for analysis of possible options. Machine Learning Techniques (MLT) help to improve decision making and can be viewed as assigning or predicting correct label based on data features (Classification Problem).

5. Experimental Results

F. Data cleaning

The fields taken and first analyzed for Missing values. Missing values can change the course and direction of a result, if not handled proper. Hence, first step is to address the missing value. Figure 5.1 depicts the output of Missing Values.

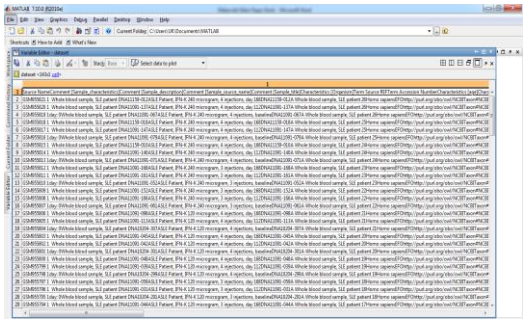


Fig 5.1. Data cleaning results

G. Decision making

The decision-making tree is one of the better-known decision-making techniques, probably due to its inherent ease in visually communicating a choice, or set of choices, along with their associated uncertainties and outcomes. Their simple structure enables use in a broad range of applications. They can be drawn by hand to help quickly outline and communicate the critical elements in a decision. Alternatively, a decision tree's simple logical structure enables it to be used to address complex multiple decision scenarios and problems with the aid of computers. The proposed work uses the CART Algorithm for forming its decision-making tree based on the criteria listed in Table 5.1.

TABLE 5.1 DECISION MAKING CRITERIA

Attribute	Decision Weight
General Attributes	1
Disease Activity	2
Symptoms	3
Test Results	4

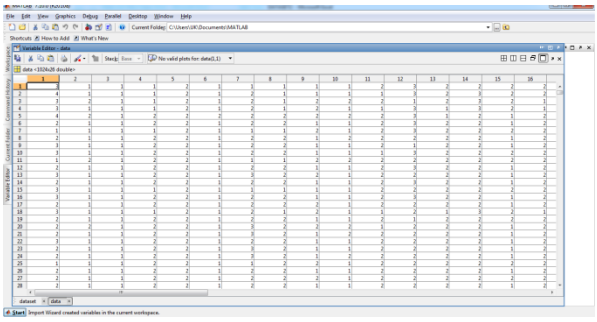


Fig 5.2 . Decision making criteria results

H. Proposed HKCART Algorithm

Input: The Healthcare dataset with n entities.

Output: A set of optimal clusters K_i.

Step 1: Likewise, any abnormal or out of range values are also not considered for preprocessing.

Step 2: Records of incorrect and missing data are not considered and assign mean mode value.

Step 3: If Multi variant attributes or more than one instance are there then

Step 4: Remove redundant value using deletion query operation

Step 5: Normalization of missing attribute instances is done by filtering.

Step 6: Processed, filtered value converted as MAT file and stored in separate database.

Step 7: Initialize dataset $\sum (F) = \{ f_1, f_2, f_3 \ldots f_n\}$ attributes.

Step 8: Identify the Outliers in the considered column $\sum (F') = \{ f_1', f_2', f_3' \ldots f_n'\}$

Step 9: Repeat, formulate the rules for identifying the similar attributes.

Step 10: do until, Identify the frequent item sets.

Step 11: Specify the threshold Mean, proportion value.

Step 12: Identify the K initial mean vector from the attributes

Step 13: Identify the distance between f_i attributes and the centroid value f_j.

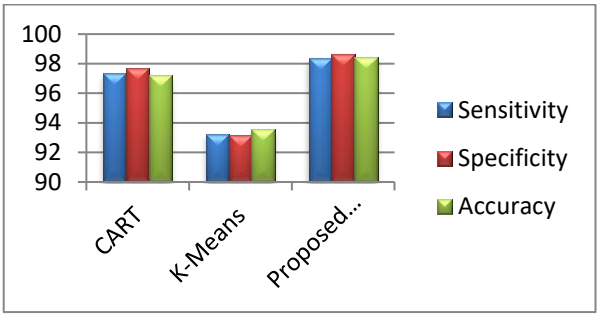
Step 14: Recalculate until new centroid f_j identified.

Step 15: Identify the end convergence.

Step 16: Fin the neighborhood active attribute rule set.

TABLE 5.2. COMPARATIVE PERFORMANCE BY DATA MINING TECHNIQUES

Algorithms	Sensitivity	Specificity	Accuracy
CART	97.33	97.66	97.19
K-Means	93.23	93.13	93.56
Proposed HKCART	98.33	98.66	98.45



6. Discussions

The management of data and utility in healthcare sector is a difficult task. But the usage of computer to store and retrieval can help. However the data generated by an individual / or a community can be enormous unless sorted out and categories can get lost in the research.

The research identifies 5 system of approach namely

1. Dataset
2. Data Cleaning
3. Missing Value Prediction
4. Dataset Preparation
5. Decision Making

Every system has its own advantage over other. Each system addresses a particular need. Putting all together one by one or at the same time, the advantages are H/c Count 98% semantic, specific and secure. The 2% error can happen in any system which is negotiable. Further study could eliminate ever this. However, H/c can't compare with Human intelligence and analytical mind can enhance the efficiency to nearly 100%. Nearly 100% is become difficult as every individual can be unique. This system approach can help to predict the disease at an early stage, so management and near cure is achievable. (Near cure be return to 100% normally is as of now impossible) but this system helps in early diagnosis and proper management. The Hospital days can be cut short and economic burden of the patent can be greatly reduced. The research work can be put to proper use in the health care management system

7. Conclusion

Computer assisted management system had pervaded through all the walks of life. The technology and associated AI help a lot in the proper healthcare management. However, managing the data accumulated and stored need to be put to proper use.

To achieve the effective management of the data and to use the same in health sector can help to reduce the hospital days the cost and enhance better management.

References

[1] CHAVES, L., and MARQUES, G. (2021) Data Mining Techniques for Early Diagnosis of Diabetes: A Comparative Study. Applied Sciences, 11(5), pp. 2218.

[2] NALLURI, S. (2020) Chronic Heart Disease Prediction Using Data Mining Techniques. Data Engineering and Communication Technology, pp. 903-912.

[3] ALDHYANI, T.H. (2020) Soft Clustering for Enhancing the Diagnosis of Chronic Diseases over Machine Learning Algorithms. Journal of healthcare engineering, 2020.

[4] MULANI, J. (2020) Deep Reinforcement Learning Based Personalized Health Recommendations. Deep Learning Techniques for Biomedical and Health Informatics, pp. 231-255.

[5] DE CNUDDE, S. (2020) A benchmarking study of classification techniques for behavioral data. International Journal of Data Science and Analytics, 9(2), pp. 131-173.

[6] SHU, X. (2020) Knowledge Discovery in the Social Sciences: A Data Mining Approach. University of California Press.

[7] JAIN, D., and SINGH, V. (2018) Feature selection and classification systems for chronic disease prediction: A review. Egyptian Informatics Journal, 19(3), pp. 179-189.

[8] SAXENA, K., and SHARMA, R. (2016) Efficient heart disease prediction system. Procedia Computer Science, 85, pp. 962-969.

[9] LI, J., LOERBROKS, A., BOSMA, H., and ANGERER, P. (2016) Work stress and cardiovascular disease: a life course perspective. Journal of occupational health, 15, pp.0326.

[10] MUSTAQEEM, A., (2017) A statistical analysis-based recommender model for heart disease patients. International journal of medical informatics, 108, pp. 134-145.

[11] SAINI, M. (2017) Prediction of heart disease severity with hybrid data mining. 2017 2nd International Conference on Telecommunication and Networks (TELNET), IEEE.

[12] AMIN, M. S. (2019) Identification of significant features and data mining techniques in predicting heart disease. Telematics and Informatics, 36, pp. 82-93.

[13] JOTHI, N., and HUSAIN, W. (2015) Data mining in healthcare–a review. Procedia Computer Science, 72, pp. 306-313.

[14] SOWMIYA, C., and SUMITRA, P. (2017). Analytical study of heart disease diagnosis using classification techniques. 2017 IEEE International Conference on Intelligent Techniques in Control, Optimization and Signal Processing (INCOS), IEEE.

[15] DEEKSHATULU, B., and CHANDRA, P. (2013) Classification of heart disease using k-nearest neighbor and genetic algorithm. Procedia Technology, 10, pp. 85- 94.

[16] BANU, M.N., and GOMATHY, B. (2014) Disease forecasting system using data mining methods. 2014 international conference on intelligent computing applications, IEEE.
[17] GANDHI, M., and SINGH, S.N. (2015) Predictions in heart disease using data mining techniques. 2015 International Conference on Futuristic Trends on Computational Analysis and Knowledge Management (ABLAZE), IEEE.
[18] SAXENA, K., and SHARMA, R. (2015) Efficient heart disease prediction system using a decision tree. International Conference on Computing, Communication & Automation, IEEE.