MAT JOURNALS

Journal of Data Mining and Management
e-ISSN: 2456-9437
Volume 3 Issue 3

A Proposed Secured Prediction System for Human Diseases Using a Genetic Algorithm Approach to Data Mining

[1]Snehal Ganesh Shinde, [2]Dr. L.M.R.J. Lobo
[1]P.G Student, [2]Associate Professor,
Department of Computer Science Engineering, WIT,
Solapur University, Solapur, Maharashtra, India
***Email:** [1]18snehalshinde@gmail.com, [2]headitwit@gmail.com*
***DOI:** http://doi.org/10.5281/zenodo.1494964*

Abstract

Many changes are happening in life styles of people in growing countries like India in recent days. Such changes in environment, diet, pollution and stress have led to the scenario that human beings are affected by microorganisms causing fatal diseases. In India, human diseases have become a major reason of deaths. A number of people have worked in this area to detect a particular disease, but it may happen that a person may be suffering from more than one disease at a time. Our attempt therefore is to detect the diseases a patient is suffering from with the use of detail symptoms given by a patient regarding his health status. This detection will enable him to have an appropriate treatment in time.In the health sector Data Mining techniques can be used to play a major role to detect a disease. The aim of the proposed system is to predict human diseases by using Association Rule Mining Algorithm name Apriori and generating optimized association rules using a Genetic Algorithm.

***Keywords:** Diseases, Association Rule Mining, Genetic Algorithms, Rules*

INTRODUCTION

Today's health-care services have come a long way to provide medical care to the patients and protect them from several diseases. Human diseases are the main reason of death throughout the world, and the larger number of deaths arises in low and middle income countries like India. Medical practitioners continuously generate large amount of data in the field of biomedical. This data can be used for the early detection of the human diseases, which can support to reduce the number of diseases. Now a day's many changes are happen in life styles of peoples in growing countries like India human diseases have become a major reason of deaths. In the health area data mining techniques play major role to discover the diseases. [5], [12]

Data Mining is a data analysis methodology used to identify hidden patterns from large amount of data. It has been successfully used in different areas for knowledge discovery. Data Mining or knowledge discovery has come out into view, as one of the most progressive areas in Communication Engineering, Information Technology and Biomedical Science. Nowadays different types of data mining methods have been used and developed. [6], [7] Data mining is an important area of research and is preferably used in Healthcare domain which is an active interdisciplinary area of research. [11]

Association rules are if / then statements that help unwrap relationships between apparently unrelated data in a relational database or other information repository. An association rule comprises of two parts, an antecedent (if) and a consequent (then). An antecedent (if) is an item found in the data. A consequent (then) is found in combination with the antecedent.

Association rules are made through breaking down information for regular if/then examples and utilizing the parameters support and confidence for distinguishing the most imperative connections.

Association Rule Mining is the way toward finding new intriguing Correlations or easygoing structures between sets of things in the exchange databases or other information storehouses. In data mining, association rule mining is an important and easy method to find frequent item sets from large dataset.

It is proposed to distinguish robust rules in databases utilizing two distinct proportions of intriguing quality. The first one is support which generates frequent item set from the provided database and the other one is confidence which is focuses on rule generation.

Frequent Item Sets- A set of attributes is termed as frequent item set if the occurrence of the
set within the database is more than a user given threshold.

Support –Support decides how frequently a given formulated rules is pertinent to a given informational index.

Confidence- Confidence decides how as often as possible things in Y show up in exchanges that contain X.

$$\text{Support, } s(X->Y) = \frac{\sigma(X \cup Y)}{N}$$

$$\text{Confidence, } c(X->Y) = \frac{\sigma(X \cup Y)}{\sigma(X)}$$

Where, X and Y disjoint item set. [10]

Genetic Algorithm (GA) depends on the standards of Genetics and Natural Selection, which has been connected in machine learning and streamlining issues and is a pursuit based enhancement method. [3]. It is usually used to discover ideal or close ideal answers for non-tractable issues which generally would take a lifetime to generate solutions. It is as often as possible used to tackle improvement issues.
Improvement is the way of toward improving something. It has an arrangement of sources and an arrangement of yields.

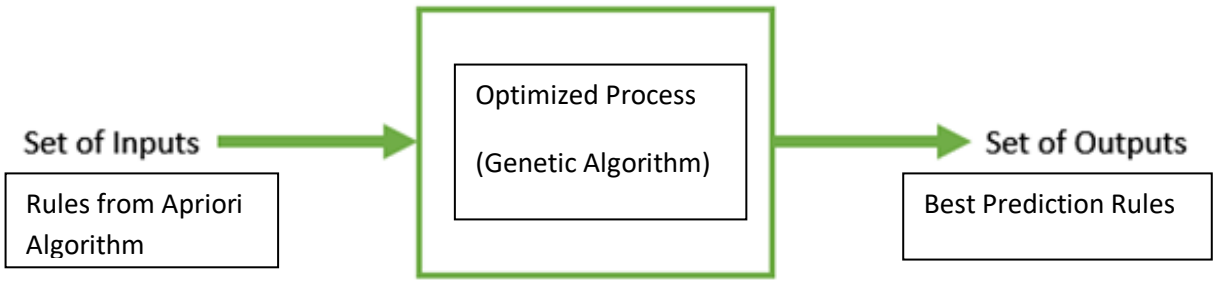

Figure: 1.*Flowchart of System Design*

Genetic Algorithms (GA) are immediate, parallel techniques for worldwide inquiry and advancement. GA is a standout amongst the most regularly utilized Evolutionary Algorithms (EA). They utilize collections with permitted number of arrangements (chromosomes), they include the gathering of parallel calculations.

It is an inquiry system utilized in registering that discovers genuine or estimated answers for advancement and hunt issues. Albeit randomized, GA's are

in no way, shape or form irregular; rather they misuse verifiable data to coordinate the hunt into the district of better execution inside the inquiry space.

The basic techniques of the GA is designed to simulate processes in natural systems that are necessary for evolution; especially those which follow the principles first laid down by Charles Darwin of survival of the fittest. The evolution generally starts from a population of randomly generated individuals and happens in generations.

In every age, the fitness of each individual chromosome in the population is assessed, numerous individuals are chosen from the present populace (in light of their fitness), and changed to shape another populace. The following emphasis of the calculation utilizes the new populace. The calculation closes when either a greatest number of parameters have been delivered, or an agreeable fitness level has been gone after the populace.

GA depends on similarity with the hereditary structure and conduct of chromosomes inside the number of inhabitants in people (chromosomes) having the establishment that people in a populace go after assets and mates. Those people which are best in every 'opposition' will create more posterity than those people that perform ineffectively. Qualities from `good' people proliferate all through the populace so two great guardians will in some cases create posterity which are superior to either parent. In this manner each progressive age will turn out to be more suited to their condition. A populace of people is kept up inside look space for a GA, every one speaking to a conceivable answer for a given issue. Every individual is coded as a limited length vector of segments, or factors, as far as letters in order, generally the parallel letter set. To proceed with the hereditary relationship these people are viewed as like chromosomes and the factors are similar to qualities. Along these lines a chromosome arrangement includes few qualities (factors). A fitness score is dispensed to every arrangement speaking to the capacities of a person to `compete'. The individual having the ideal (or for the most part close ideal) fitness score is looked for. The main ingredients of GA are Chromosomes, Selection, Recombination and Mutation.

Selection

An extent of the current populace is chosen to breed another age amid each progressive age. Fitness based process is utilized to choose singular arrangements where fitter arrangements (as estimated by a wellness work) are regularly more prone to be chosen. In this stage elitism could be utilized – the best n people are straightforwardly exchanged to the people to come. The elitism guarantees, that the estimation of the improvement work can't get most exceedingly bad (once the extremis is achieved it would be kept).

Crossover

The most widely recognized compose is single point hybrid. In single point hybrid, we pick a locus time when you swap the rest of the alleles from one parent to the next. The kids take one segment of the chromosome from every individual parent. Chromosome is broken dependent on the arbitrarily chose hybrid point. This specific technique is called single point hybrid on the grounds that here just a single hybrid point exists. Some of the time just a single youngster is made, however by and large both posterity are made and put into the new populace. Hybrid does not generally happen. Once in a while, in light of a set likelihood, no hybrid happens and the guardians are straightforwardly duplicated to the new populace.

Mutation
We have another populace loaded with people (Chromosomes) where some are straightforwardly duplicated, and others are created by hybridisation. With the end goal to guarantee that the people are not all the very same, we permit a little possibility of change. We experience every one of the alleles of the considerable number of people, and if that allele is chosen for transformation, we either transform it by a little sum or supplant it with esteem. Change is genuinely basic. Notwithstanding, Mutation is crucial to guaranteeing hereditary assorted variety inside the populace. Genetic Algorithm is a randomized calculation that could be kept running for quite a while to acquire an ideal arrangement. [3]

Association Rule Mining is the most ground-breaking procedure in information digging for producing rules. Apriori is most critical calculation based approach for producing association rules. We have therefore utilized Apriori Algorithm to create solid and legitimate rules. These standards are then advanced utilizing Genetic Algorithm to get best principles.

The present proposed work includes:
The input to be used is person / patient details (dataset). The implementation good include using Arrhythmia (Heart Disease and Abnormal Heart Rhythm) disease dataset available in UCI repository.

1. Performing association rule mining on dataset for generating required association rules.
2. Optimize association rules using Genetic Algorithm to get the best rules.

Related Work
Rahman et al. [1] Designed and executed a specialist framework dependent on continuous patient criticism that expects to give a reasonable choice, which would help in customizing the administration and additionally recognizing and distinguishing of any shunt glitches without the need to contact or visit the doctor's facility, for hydrocephalus administration and shunt conclusion. The win-prolog programming environment was used for developing the patient feedback expert system. The patient's details and symptoms are inputs, and the result of patient feedback analysis system is either that the patient needs to contact the physician or the problem is handled by modifying the opening times of the valve or assure him that the cause of the symptoms is not due to shunt complication. The system has the ability to identify the shunt state, i.e. problem existing or not and if yes identify such problem.

Kaysi et.al. [2] Developed an investigation that planned to foresee which patients become more cheerful and insight because of tDCS treatment by examining electroencephalography (EEG) of MDD patients that was gathered toward the beginning of tDCS treatment. This was accomplished through ordering power otherworldly thickness (PSD) of resting-state EEG utilizing bolster vector machine (SVM), straight segregate examination (LDA) and outrageous learning machine (ELM). Members were named as enhanced or not enhanced dependent on the adjustment in state of mind and subjective scores. The obtained classification results of all channel pair combinations were used to identify the most relevant brain regions and channels for this classification task. This represented an encouraging sign that EEG-based classification may help to tailor the selection of patients for treatment with tDCS brain stimulation.

Syam et.al. [3] Used Medical images for retrieval and the feature extraction along with colour, shape and texture feature extraction to extract the query image from the database medical images. At the point when a question picture was given, the

highlights were separated and after that the Genetic Algorithm-based closeness measure was performed between the inquiry picture highlights and the database picture highlights. The Squared Euclidean Distance (SED) registered the comparability measure in deciding the Genetic Algorithm fitness. Consequently, from the Genetic Algorithm-based similitude measure, the database pictures that were pertinent to the given question picture were recovered. The CBIR procedure was assessed by questioning distinctive restorative pictures and the recovery effectiveness was assessed in the recovery results.

Dragulescu and Albu [4] developed an expert system to make some predictions regarding the hepatitis infection. This system implementation used three algorithms, Bayes's theorem, Aitken's formula and Logistic model
The system presented three important parts:
-The First part was a logical inference which was used to decide what type of hepatitis virus present for a new patient. The possibilities were B, B+D and C.
-The second part of the system was used to see the type and the grade using methods from statistical inference.
-The third part of the system - is made for the patients infected with hepatitis C virus and it predicted the biological parameters evolution during the treatment using artificial neural networks.

Singh et.al. [5] Developed a framework based on associative classification techniques on heart dataset for early diagnosis of heart based diseases. The attributes considered related to cause of heart diseases were - gender, age, chest pain type, blood pressure, blood sugar. The used Data mining algorithms namely Apriori, FP-Growth, Naive bayes, ZeroR, OneR, J48 and k-nearest neighbour. On basis of best results, using hybrid technique for classification associative they achieved a prediction accuracy of 99.19%

Nahar et.al. [6] Developed a research which assessed the performance of six well-known classification algorithms: Naive Bayes, SMO, IBK, AdaBoostM1, J48 and PART, using a number of performance matrices. The research explored medical (domain knowledge) knowledge based feature selection (MFS) on a real-life dataset and noted performance improvements compared to computer-automated feature selection for majority of the considered techniques and across majority of the performance measures. The findings could assist the design of a heart disease CAD and might also guide exploring the complicated symptoms, risk and prevention factors of the different disease for particular group of population.

Saraee et.al. [7] Proposed an approach for using data mining in classifying mortality rate related to accidents in children under 15. These data were gathered from the patient files which were recorded in themedical record section of the Alzahra Hospital in Isfahan.

Ranganatha et.al. [8] Developed a task to store restorative data of patients who sought hospitalization for coronary illness and calculations were kept running on that data and result were given as client reasonable words and chart. Data mining algorithms ID3 and Naïve Bayesian were used. The user had to login to enter the patient information. After login, patient information page is displayed where the user fills patient history form and it is given as an input to the algorithms. The algorithms are executed to give the result in the form of decision tree in case of ID3 and probability in case of Naïve Bayesian. The attributes of heart disease dataset included the Name, Age, Gender, Chest

Pain Type, Rest ECG, CA (Coronary Angioplasty), Exang, Slope, FBS (Fasting Blood Sugar).

Palaniappan and Awang [9] built up an Intelligent Heart Disease Prediction System utilizing information mining strategies. Information mining procedures are Decision Trees, Naïve Bayes and Neural Network. It could answer complex questions for diagnosing coronary illness. To assemble the mining models the CRISP-DM technique were utilized. Business understanding, information understanding, information arrangement, displaying, assessment, sending are six noteworthy stages. To construct and access the models DMX question dialect and capacities are utilized. Lift Chart and Classification Matrix techniques were utilized to assess the adequacy of the models.

Song et.al. [10] Developed a project which establishes the intelligent diagnosis model of lung cancer based on the testing data from clinical diagnoses and by means of interactive regulation tapping. They respectively analysed 12 targets of 100 patients through multi-subjects combined means of radioimmunology, enzyme linked immunosorbent assay and chemistry so as to discover the interactive regulation between cancer and its likely causes and guide the diagnosis and prevention of lung cancer with the regulation model. The result showed that this method is superior to the conventional statistics of quantitative medicine.

Pandey [11] built up an investigation that intended to give a survey of information mining in the domain of human services. They talked about that information mining can be gainful in therapeutic space. Focal points and inconveniences of every now and again utilized information mining systems in the area of human services and therapeutic information have been looked at. Distinctive information mining methods, their points of interest and disadvantages are investigated. For compelling usage of these systems in human services space, there was a need to improve and secure wellbeing information sharing among different gatherings. Uniqueness of information mining as for restorative information is likewise tended to. The requirements and troubles identified with protection affectability and vast volume of medicinal information assume imperative job in determination of the specific information mining procedure. Moral and legitimate parts of medicinal information were additionally essential viewpoints. In view of its appropriateness to all individuals restorative information can have an extraordinary status.

Xao [12] built up a forecast model of coming year's FBG, in view of four years' verifiable restorative examination information, utilizing customary information mining procedures with a novel calculation to appraise the FBG change likelihood and a proposed highlight determination calculation, that joins the component significance scores of gathering learning and Sequential Backward Selection (SBS) calculation to choose an ideal element subset. Exploratory information was gathered from a medicinal examination database containing 108,386 clients, in which 7,136 individuals have four years' records. By contrasting the exploratory outcomes and arbitrary woodland and SVM, the element determination technique can adequately enhance the model execution.

Methodology

An Execution of Proposed System

The below figure: 2 show an execution of proposed system of our proposed work. In this system the patients' disease details are collected and stored in a dataset. This is given to an association rule mining algorithm name Apriori. The algorithm

returns rules that represent the dataset. However these rules are too many and not optimized. Genetic algorithm is then applied to generate the best rules that predict the disease suffered by the patient.

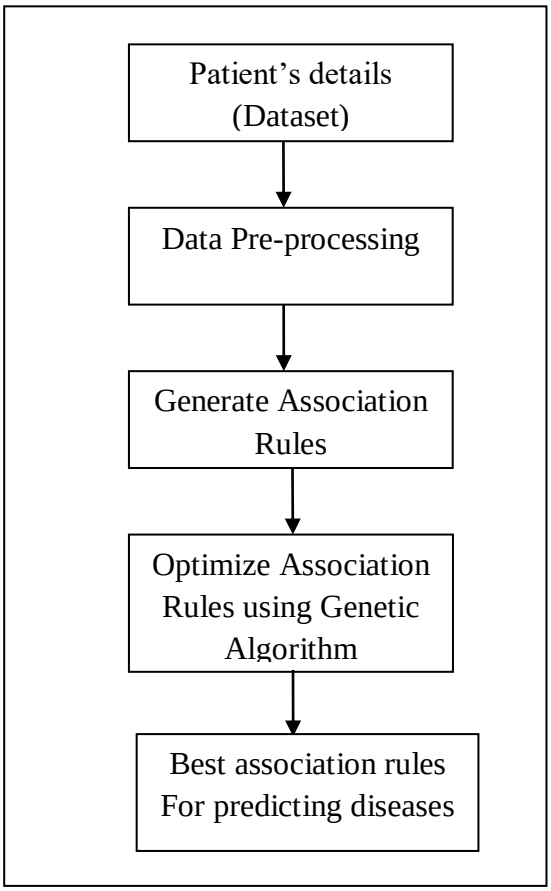

Figure: 2. *An Execution of Proposed System*

EXPECTED RESULT

The anticipated yield would be the best affiliation rules created by applying hereditary calculation to advance the standard delivered by Apriori calculation of affiliation rule mining. These tenets will be utilized for the ideal expectation of sicknesses.

CONCLUSION

In this system Association Rule Mining Algorithm will be used for generating the rules from the disease dataset. These rules are then optimized using Genetic Algorithms to get the best rules that represents the best prediction. We prefer to use Genetic algorithms since a problem is NP hard and can be solved by Genetic Algorithms in generations. This system can be used in the medical sector. It is observed that many people lose their lives due to untimely detection of diseases. Such a proposed system would detect a disease in time and required treatment can be given to save the patient. Thus it has a social cause.

REFERENCES

1. AbdelRahmanAlkharabsheh, LinaMo mani,Nayel Al-Zu'bi,Waleed Al-Nuaimy"
   An expert system for hydrocephalus pa tient feedback" in Annual International IEEE Conference of the IEEE Engineering in Medicine and Biology,2010
2. Alaa M. Al-Kaysi, Ahmed Al-Ani, Colleen K. Loo, Michael

Breakspear,Tjeerd W. Boonstra "Predicting brain stimulation treatment outcomes of depressed patientsthroug h the classification of EEG oscillations " in proceedings of 38th Annual International IEEE Conference of the Engineering in Medicine and Biology Society (EMBC),2016

3. B. Syam, J. Sharon Rose Victor, Y. SrinivasaRao"Efficient similarity measure via Genetic algorithm for content basedmedical image retrieval with extensive features" in proceedings of IEEE International Mutli-Conference on Automation, Computing, Communication, Control and Compressed Sensing (iMac4s),2013
4. DoinaDragulescu, Adriana Albu "Expert System for Medical Prediction s" in proceedings of 4th International IEEE Symposium on Applied Computational Intelligence and Informatics,2007
5. Jagdeep Singh, AmitKamra and Harbhag Singh "Prediction of heart diseases using associative classification" in proceedings of 5th International IEEE Conference on Wireless Networks and Embedded Systems (WECON), 2016
6. JesminNahar, Tasadduq Imam, Kevin S. Tickle, Debora Garcia-Alonso "Medical Knowledge based Data Mining forCardiac Stress Test Diagnostics"in proceeding of 2nd Asia-Pacific World Congress on Computer Science and Engineering (APWC on CSE),2015
7. MohammadHosseinSaraee, ZahraEhgh aghi,HodaMeamarzadeh,BahareZibane zhad"Applying data mining in medical data with focus on mortality related to accident in children" in proceedings of International IEEE Multitopic Conference,2008
8. S.Ranganatha, H.R.Pooja Raj, C.Anusha, S.K. Vinay "Medical data mining and analysis for heart disease dataset using classification technique" in proceedings of National IEEE Conference on Challenges in Research & Technology in the Coming Decades (CRT 2013), 2013
9. SellappanPalaniappan,RafiahAwang"I ntelligent heart disease prediction syste m using data mining techniques" in proceedings of International IEEE/ACS Conference on Computer Systems and Applications,2008
10. Shaoyun Song, Yu Ma"The Research and Application of Technology in the Diagnosis of Lung Cancer Warning Association Rule Mining" in proceedings of 8th International Conference on Information Technology in Medicine and Education (ITME),2016
11. Subhash Chandra Pandey "Data mining techniques for medical data: A review" in proceedings of International IEEE Conference on Signal Processing, Communication, Power and Embedded System (SCOPES),2016
12. WenxiangXao, Fengjing Shao, Jun Ji, Rencheng Sun, ChunxiaoXing"Fasting Blood GlucoseChangePredictionModelBased on MedicalExamination Data and Data Mining Technique" in proceedings ofInternational IEEE Conference on Smart City/SocialCom/SustainCom (SmartCity),2015

Cite this article as: Snehal Ganesh Shinde, & Dr. L.M.R.J. Lobo. (2018). A Proposed Secured Prediction System for Human Diseases Using a Genetic Algorithm Approach to Data Mining. Journal of Data Mining and Management, 3(3), 25–32. http://doi.org/10.5281/zenodo.1494964