AGGREGATED K –MEANS CLUSTERING AND GENETIC ALGORITHM FOR PREDICTING HEART DISEASES

[1]Rajkamal.j
Assistant Professor

[2]Suruthi S
Sruthimaniapr4@gmail.com

[3]Swathi B
swathibabucse11@gmail.com

[2][3]UG students
Department of Computer Science and Engineering
T.J.S. Engineering College

Abstract

Heart disease is one of the main sources of demise around the world and it is imperative to predict the disease at a premature phase. The computer aided systems help the doctor as a tool for predicting and diagnosing heart disease. The objective of this project is to widespread about Heart related cardiovascular disease and to brief about existing decision support systems for the prediction and diagnosis of heart disease supported by data mining and hybrid intelligent techniques .

Index Terms – ***Community detection, Super imposed communities, Seed dispersion, Personalized Page Rank***.

1.Introduction

Data mining is the computational process of discovering patterns in large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems. It is an interdisciplinary subfield of computer science. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use. Aside from the raw analysis step, it involves database and data management aspects,
preprocessing, model and inference considerations,
interestingness metrics, complexity considerations, post-processing of discovered structures, visualization, and online updating. Data mining is the analysis step of the "knowledge discovery in databases" process, or KDD.
The term is a misnomer, because the goal is the extraction of patterns and knowledge from large amounts of data, not the extraction (*mining*) of data itself.[5] It also is a buzzword[6] and is frequently applied to any form of large-scale data or information processing (collection, extraction, warehousing, analysis, and statistics) as well as any application of computer decision support system, including artificial intelligence, machine learning, and business intelligence. The book *Data mining: Practical machine learning tools and techniques with Java*[7] (which covers mostly machine learning material) was originally to be named just *Practical machine learning*, and the term *data mining* was only added for marketing reasons.[8] Often the more general terms (*large scale*) *data analysis* and *analytics* – or, when referring to actual methods, *artificial intelligence* and *machine learning* – are more appropriate.
The actual data mining task is the automatic or semi-automatic analysis of large quantities of data to extract previously unknown, interesting patterns such as groups of data records (cluster analysis), unusual records (anomaly detection), and dependencies (association rule mining, sequential pattern mining). This usually involves using database techniques such as spatial indices. These patterns can then be seen as a kind of summary of the input data, and may be used in further analysis or, for example, in machine learning and predictive analytics. For example, the data mining step might identify multiple groups in the data, which can then be used to obtain more accurate prediction results by a decision support system. Neither the data collection, data preparation, nor result interpretation and reporting is part of the data mining step, but do belong to the overall KDD process as additional steps.
The related terms *data dredging*, *data fishing*, and *data snooping* refer to the use of data mining methods to sample parts of a larger population data set that are (or may be) too small for reliable statistical inferences to be made about the validity of any patterns discovered. These methods can, however, be used in creating new hypotheses to test against the larger data populations.
Heart disease is one of the main sources of demise around the world and it is imperative to predict the disease at a premature phase. The computer aided systems help the doctor as a tool for predicting and diagnosing heart disease. The objective of this project is to widespread about Heart related cardiovascular

disease and to brief about existing decision support systems for the prediction and diagnosis of heart disease supported by data mining and hybrid intelligent techniques .

1.2 PROJECT PLATFORM SPECIFICATION

STAND-ALONE:

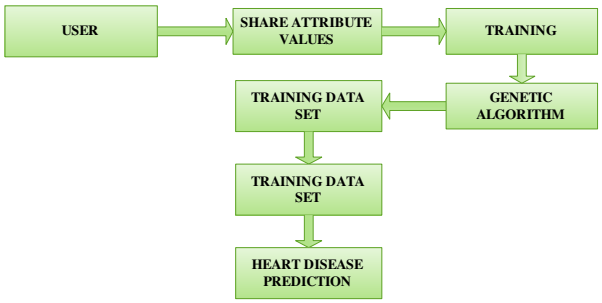
JAVA:

Java is a programming language originally developed by James Gosling at Sun Microsystems (which is now a subsidiary of Oracle Corporation) and released in 1995 as a core component of Sun Microsystems' Java platform. The language derives much of its syntax from C and C++ but has a simpler object model and fewer low-level facilities. Java applications are typically compiled to byte code (class file) that can run on any Java Virtual Machine (JVM) regardless of computer architecture. Java is general-purpose, concurrent, class-based, and object-oriented, and is specifically designed to have as few implementation dependencies as possible. It is intended to let application developers "write once, run anywhere". Java is considered by many as one of the most influential programming languages of the 20th century, and widely used from application software to web application.

One characteristic of Java is portability, which means that computer programs written in the Java language must run similarly on any supported hardware/operating-system platform. This is achieved by compiling the Java language code to an intermediate representation called Java byte code, instead of directly to platform-specific machine code. Java byte code instructions are analogous to machine code, but are intended to be interpreted by a virtual machine written specifically for the host hardware. End-users commonly use a Java Runtime Environment (JRE) installed on their own machine for standalone Java applications, or in a Web browser for Java applets. Standardized libraries provide a generic way to access host-specific features such as graphics, threading and networking. A major benefit of using byte code is porting. However, the overhead of interpretation means that interpreted programs almost always run more slowly than programs compiled to native executables would, and Java suffered a reputation for poor performance. This gap has been narrowed by a number of optimization techniques introduced in the more recent JVM implementations.

Java uses an automatic garbage collector to manage memory in the object lifecycle. The programmer determines when objects are created, and the Java runtime is responsible for recovering the memory once objects are no longer in use. Once no references to an object remain, the unreachable memory becomes eligible to be freed automatically by the garbage collector. Something similar to a memory leak may still occur if a programmer's code holds a reference to an object that is no longer needed, typically when objects that are no longer needed are stored in containers that are still in use. If methods for a nonexistent object are called, a "null pointer exception" is thrown.

SYSTEM ARCHITECTURE:



2. TESTING:

SYSTEM TESTING:

After finishing the development of any computer based system the next complicated time consuming process is system testing. During the time of testing only the development company can know that, how far the user requirements have been met out, and so on.

Following are the some of the testing methods applied to this effective project:

SOURCE CODE TESTING:

This examines the logic of the system. If we are getting the output that is required by the user, then we can say that the logic is perfect.

SPECIFICATION TESTING:

We can set with, what program should do and how it should perform under various condition. This testing is a comparative study of evolution of system performance and system requirements.

MODULE LEVEL TESTING:

In this the error will be found at each individual module, it encourages the programmer to find and rectify the errors without affecting the other modules.

UNIT TESTING:

Unit testing focuses on verifying the effort on the smallest unit of software-module. The local data structure is examined to ensure that the date stored temporarily maintains its integrity during all steps in the algorithm's execution. Boundary conditions are tested to ensure that the module operates properly at boundaries established to limit or restrict processing.

INTEGRATION TESTING:

Data can be tested across an interface. One module can have an inadvertent, adverse effect on the other. Integration testing is a systematic technique for

constructing a program structure while conducting tests to uncover errors associated with interring.

VALIDATION TESTING:
It begins after the integration testing is successfully assembled. Validation succeeds when the software functions in a manner that can be reasonably accepted by the client. In this the majority of the validation is done during the data entry operation where there is a maximum possibility of entering wrong data. Other validation will be performed in all process where correct details and data should be entered to get the required results.

RECOVERY TESTING:
Recovery Testing is a system that forces the software to fail in variety of ways and verifies that the recovery is properly performed. If recovery is automatic, re-initialization, and data recovery are each evaluated for correctness.

SECURITY TESTING:
Security testing attempts to verify that protection mechanism built into system will in fact protect it from improper penetration. The tester may attempt to acquire password through external clerical means, may attack the system with custom software design to break down any defenses to others, and may purposely cause errors.

PERFORMANCE TESTING:
Performance Testing is used to test runtime performance of software within the context of an integrated system. Performance test are often coupled with stress testing and require both software instrumentation.

BLACKBOX TESTING:
Black- box testing focuses on functional requirement of software. It enables to derive ets of input conditions that will fully exercise all functional requirements for a program.
Black box testing attempts to find error in the following category:

- Incorrect or missing function
- Interface errors
  - Errors in data structures or external database access and performance errors.

.3.LITERATURE REVIEW:

Prediction of heart disease using a hybrid technique in data mining classification author proposed that Heart disease prediction is treated as most complicated task in the field of medical sciences. Thus there arises a need to develop a decision support system for detecting heart disease of a patient. In this paper, we propose efficient genetic algorithm hybrid with the back propagation technique approach for heart disease prediction. Today medical field have come a long way to treat patients with various kind of diseases. Among the most threatening one is the Heart disease which cannot be observed with a naked eye and comes instantly when its limitations are reached. Bad clinical decisions would cause death of a patient which cannot be afforded by any hospital. To achieve a correct and cost effective treatment computer-based and support Systems can be developed to make good decision. Many hospitals use hospital information systems to manage their healthcare or patient data. These systems produce huge amounts of data in the form of images, text, charts and numbers. Sadly, this data is rarely used to support the medical decision making. There is a bulk of hidden information in this data that is not yet explored which give rise to an important query of how to make useful information out of the data. So there is necessity of creating an excellent project which will help practitioners predict the heart disease before it occurs. The main objective of this paper is to develop a prototype which can determine and extract unknown knowledge (patterns and relations) related with heart disease from a past heart disease database record. It can solve complicated queries for detecting heart disease and thus assist medical practitioners to make smart clinical decisions which traditional decision support systems were not able to. By providing efficient treatments, it can help to reduce costs of treatment.

4.EXISTING SYSTEM:

Many hospitals manage healthcare data using healthcare information system; as the system contains huge amount of data, used to extract hidden information for making intelligent medical diagnosis. To develop this system, medical terms such as cholesterol like 13 input attributes are used.

5. PROPOSED SYSTEM:

The proposed work is to predict the heart disease from the input attributessuch as Age,sex, Chest pain type, Resting blood pressure, Serum cholesterol, resting electrographic results, Fasting blood sugar, Maximum heart rate achieved, Exercise induced agina, ST depression induced by exercise relative to rest, Slope of the peak exercise ST segment.

5.1 MODULES:

Preprocessing:

Preprocessing is the process of cleaning the data's in order to save the time and also to prevent the data error. As data is in free format and extracted from web, lot of irrelevant information's like HTML tags,

special characters, spelling mistakes, removed from document to make data easy for further use.

Genetic Algorithm
In Artificial Intelligence, Genetic Algorithm is a search technique which uses the process of natural selection. Genetic algorithms provide solutions to optimization and search problems by using techniques such as inheritance, mutation, selection, and crossover. A typical genetic algorithm requires genetic representation of the solution domain and a fitness function to evaluate the solution domain.

Neural Network
Neural networks (also referred to as connectionist systems) are a computational approach, which is based on a large collection of *neural* units (AKA artificial neurons), loosely modeling the way a biological brain solves problems with large clusters of biological neurons connected by axons.

6.ALGORITHM:

1. Genetic Algorithm

In Artificial Intelligence, Genetic Algorithm is a search technique which uses the process of natural selection. Genetic algorithms provide solutions to optimization and search problems by using techniques such as inheritance, mutation, selection, and crossover. A typical genetic algorithm requires genetic representation of the solution domain and a fitness function to evaluate the solution domain. A **genetic algorithm** (GA) is a method for solving both constrained and unconstrained optimization problems based on a natural selection process that mimics biological evolution. The **algorithm** repeatedly modifies a population of individual solutions. **Mutation** is a **genetic** operator used to maintain **genetic** diversity from one generation of a population of **genetic algorithm** chromosomes to the next. It is analogous to biological **mutation**. **Mutation** alters one or more gene values in a chromosome from its initial state. A **fitness function** is a particular type of objective **function** that is used to summarise, as a single figure of merit, how close a given design solution is to achieving the set aims. In artificial intelligence, **genetic programming** (GP) is a technique whereby computer programs are encoded as a set of **genes** that are then modified (evolved) using an evolutionary algorithm (often a **genetic** algorithm - "GA"). The result is a computer **program** able to perform well in a predefined task. **Fuzzy logic** is a form of many-valued **logic** that deals with approximate, rather than fixed and exact reasoning. Compared to traditional binary **logic** (where variables may take on true or false values), **fuzzy logic** variables may have a truth value.

CONCLUSION:
In this paper, Heart disease is one of the main sources of demise around the world and it is imperative to predict the disease at a premature phase. The computer aided systems help the doctor as a tool for predicting and diagnosing heart disease. The objective of this project is to widespread about Heart related cardiovascular disease and to brief about existing decision support systems for the prediction and diagnosis of heart disease supported by data mining and hybrid intelligent techniques

REFERENCES:

[1] D. Ratnam, P. HimaBindu, V. MallikSai, S. P. Rama Devi and P. RaghavendraRao, "Computer-Based Clinical Decision Support System for Prediction of Heart Diseases Using Naïve BayesAlgorithm", International Journal of Computer Science and Information Technologies, vol. 5, no. 2 (2014) pp.2384-2388
. [2] T. Santhanam and E. P. Ephzibah, "Heart Disease Prediction Using Hybrid Genetic Fuzzy Model", Indian Journal of Science and Technology, vol.8, no. 9, (2015), pp.797–803.
[3] G. Purusothaman and P. Krishnakumari, "A Survey of Data Mining Techniques on Risk Prediction: Heart Disease",Indian Journal of Science and Technology, vol. 8, no. 12, (2015).
[4] K. Srinivas, G. RaghavendraRao and A. Govardhan, "Analysis of Coronary Heart Disease and Prediction of Heart Attack in Coal Mining Regions Using Data Mining Techniques", Proceedings of 5th International Conference on Computer Science & Education, China, (2010), pp. 24–27
. [5] K. Thenmozhi and P. Deepika, "Heart Disease Prediction Using Classification with Different Decision TreeTechniques", International Journal of Engineering Research and General Science, vol. 2, no. 6, (2014), pp.6-11. [6] S. Chaitrali, Dangare and S. Apte, "Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques", International Journal of Computer Applications, vol. 47, no. 10, (2012), pp. 0975 –88.
[7] J Peter and K. Somasundaram, "An Empirical Study on Prediction of Heart Disease Using Classification DataMining Techniques", Proceedings of IEEE International Conference on Advances In Engineering, Science And Management (ICAESM), (2012), pp. 514-518.
[8] H. D. Masethe and M. A. Masethe, "Prediction of Heart Disease using Classification Algorithms", Proceedings of the World Congress on Engineering and Computer Science(WCECS), San Francisco, USA.