International Journal of Scientific Engineering and Research (IJSER)
www.ijser.in
ISSN (Online): 2347-3878, Impact Factor (2014): 3.05

A Trivial Survey on Preset Prediction of Coronary Heart Disease Using Data Mining and Soft Computing Technique

Sowmya N.[1], Dr. R .Vijayabhanu[2]

Abstract: ***Knowledge Discovery in Databases is a widely used in miscellaneous areas. The designation of data mining methodology has been conveyed through such relevance, integrating it with other technologies such as soft computing. The consequence on the prediction of coronary heart disease has been achieved through four –phases proposal. This proposal has been coiled with the tactics of classification and optimization. Coalesce of such methodologies makes the data mining system more rapid and more consistent. A cavernous survey of the literature is done to be evidence for the various purposes and achievements of soft computing methodologies along with data mining.***

Keywords: Coronary Heart Disease, Data mining, Decision tree, Fuzzy system, optimization

1. Introduction

Today's world of preset data anthology and futuristic database provides us with a copious amount of information in various e-formats. Detection and prediction with certain knowledge has become effective through data analytics. Extraction of acquaintance and preprocessing of missing attributes increase the process of such system. Splitting of huge database into branches with the classifiers such as Decision tree enhance the process of prediction. The fuzziness of the system should be impassive and optimized to gain better accuracy of the prophecy. The significance of the prediction in the health check domain is rapidly escalating now -a -days. One of such medical forecast that prevents heart attacks and sudden deaths is the prediction of Coronary Heart Disease (CHD). This manuscript deals with prophecy of the disease in the following sections: **Section I** - Preprocessing of the heart disease database to salvage the efficient attributes, **Section II**- Classification of the colossal database into prediction nodes, **Section III** –Eliminate the vagueness of the system and **Section IV**- Optimize to obtain the accuracy. The Survey on this relevance presents a better way of accepting a proposal by incorporating the above sections.

2. Data Mining and Soft Computing Techniques

2.1 Data Mining

The digital world of today produces a copious source of data known as big data. This big data consists of non-trivial, hidden, previously unknown and potentially valuable data. Such data can be used effectively for the prediction of future trends. This can be done through data mining processes. The veiled patterns and relationships can be retrieved knowledge based on mining. This effective tool is used to identify the acquaintance based on convinced criteria. Thus, it is sometimes known as data or knowledge discovery [14].

The ultimate goal of data archeology is to extract useful data and analyze them with different perceptions and gather them into useful information. There are large database of information that has been stored in various electronic forms which may consist of curtailed, noisy and inconsistent data. The data mining methodologies are used to harvest the embedded information which is used as a cause of knowledge for decision building.

The electronic data are preprocessed by the data or pattern analysis to construct the predictive modules. These modules are rooted with the algorithms such as K-means, SVM, KNN, CART, Naïve Bayes, etc to envisage the indispensable knowledge or information for decision making.

2.1.1 Applications of Data Mining:

Data Mining as considered as a powerful tool which is capable of conducting decision making and for forecasting future trends of market. Data Mining tools and techniques can be successfully applied in various fields in various forms [13]. The medical and healthcare data mining are remarkable, since that there are huge and intricate volumes of data that have been generated by various un-automated analysis and healthcare activities.

The following are considered as the major issues of medical data mining,

a) **Heterogeneity of Medical Data-** The medical data are in huge volumes and they pursue complexity by nature. The diverse interpretation of the physicians and the canonical form of the data also adds the heterogeneity of the data.
b) **Ethical, Legal and Social Issues-** The ownership of the data and the administrative issues are considered to enhance the privacy and the security of the human data.

2.1.2 Decision Tree

A decision tree is a widely used data mining classifier, which incorporates both nominal and numerical data. Being uttered as a recursive partition of the instance space, the decision trees use certain discrete function of the input attributes. According to a Survey on Decision Tree Algorithm for Classification by Brijain R Patel *et al.* (2014) [4], to extort models from a large data set there are two forms of data analysis namely classification and prediction. Such analysis can be effectively worn to foresee for prospecting data trends.

The decision trees are the well known paradigm to depiction any discrete value classifier that is proficient of handling datasets that may have error and missing values. Consequently, these trendy approaches are used to predict the accuracy of CHD in a choice of related studies. Widely used heart disease datasets in decision tree research consists of 303 numbers of instances and 75 numbers of attributes.

'The tree complexity has a decisive effect on its accuracy' was the statement by Breiman *et al.* (1984) [3]. This tree complexity of the decision tree is clearly achieved by using stopping criteria and the pruning methodologies. Thus, the decision tree inducers provide exclusive potential to boost the conventional statistical forms of analysis.

The decision tree inducers are the algorithms that involuntarily construct a decision tree from a specified dataset. The primary objective is to obtain the optimal decision tree thereby minimizing the generalization error. The decision tree inducers can be reflected on either of top-down or bottom-up approaches. The greedy algorithm is considered to the indispensable learning approach that proceeds with the recursive top down approach of decision tree structure. The decision tree algorithm has experienced a lot in the world of data mining. These inducers algorithms such as CART, C4.5, and C5 are largely used in the predictions.

CART: Breiman *et al.*, (1984) [3] has projected the classification algorithm called the Classification and regression tree (CART) for constructing binary trees in which each internal node precisely has two retiring edges. The CART algorithm has been also termed as Hierarchical Optimal Discriminate Analysis (HODA) that enables the users by providing the prior probability distribution. The cost-complexity Pruning and Gini index are used to prune the tree obtained from the CART algorithm and as the impurity measure for selecting attribute respectively. This algorithm makes use of both categorical and numeric variables either to construct classification or regression trees and thus it is a non-parametric decision tree learning technique. *Persi Pamela et al*. used Classification and Regression Tree (CART) algorithm along with the Particle Swarm Optimization (PSO) to predict 94% accuracy of CHD [16]. To envisage 87.74 % accuracy by means of 19 attributes Dursan *et al.* has used C5 and CART decision trees [7].

C4.5: The algorithm proposed by Ross Quinlan is considered to be an extension of Quinlan's earlier ID3 algorithm has wisely used to generate a decision tree. C4.5 has been used for classification rather than regression, which are regularly referred to as a statistical classifier. Splitting criteria is used as the information gain in this algorithm. Both categorical and numerical values data are admitted even with missing values, thus by increasing its gain calculation. C4.5 algorithm provides an expedient way to lever the continuous values by generating threshold and dividing the attributes across the threshold value.

2.2 Fuzzy Systems

A fuzzy system is considered to be the conservatory of the traditional fuzzy mathematics. The fundamental of the fuzzy mathematics are laid by the fuzzy sets and the fuzzy logic. The multi-valued logic that allows transitional values to be defined between conventional evaluations like 0/1, true/false, yes/no, high/low, etc are known as Fuzzy Logic (FL) [20]. The fuzzy set is considered to be the basic notion of the fuzzy system. Membership functions of fuzzy sets can be distinct in any integer of ways as long as they follow the rules of the description of a fuzzy set.

The fuzzy logic is said to be the superset of Boolean logic that has been unmitigated to grip the concept of the partial certainty values between 'completely true' and 'completely false'. The fuzzy system logic recognizes further than simple true and false values. The expertise considers fuzzy logic as "a constitution of knowledge depiction appropriate for notions that cannot be defined accurately, but which depend upon their context".

The Classical set or the Crisp set contains the objects that can convince accurate properties of membership. The Crisp membership functions have values of either one or zero. But the fuzzy set contains the vague properties of membership in correspondence to their objects. Fuzzy is said to determine "possibility" rather than "probability". The impetus of the fuzzy logic is to alleviate difficulties in developing and analyzing complex systems encountered by the conventional mathematical utensils [2].

Fuzzy Logic Process

The FL process is a progression of computing, reasoning and modeling with the fuzzy familiarity. Despite the fact that the massiveness of the information we incorporate each day with fuzzy, most of the actions or decisions implemented by humans or machines are crisp or binary. Fuzzy logic provides a substitute way to represent linguistic and subjective attributes of the real world in computing. It is able to be applied to control systems and other applications in order to improve the efficiency and simplicity of the design process.

The ultimate scenario of fuzzy system is the prospect for modeling of circumstances which are inherent and simultaneous numerical and linguistic data. Fuzzy systems are extensively used for modeling, simulating and replicating many genuine tribulations.

The figure 1 illustrates the structure of fuzzy system.

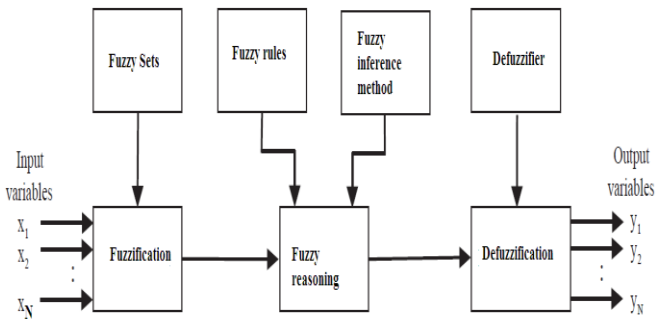

Figure 1: Structure fuzzy system

- **Fuzzification -** Converts the crisp input to a linguistic variable using the membership functions stored in the fuzzy data base.

The conversion of real inputs to fuzzy set values is the preliminary of the fuzzy system. In the real world, hardware and manuals generates crisp data, but these data are subject to investigational errors. By establishing the fact base of the fuzzy system we identify the input and output of the system. The IF THEN rules are coiled and uses unprocessed data to develop a membership function.

- **Fuzzy Inference System**

The Fuzzy rules are based on fuzzy premises and fuzzy consequences. Truth value for the premise of each rule is computed, and applied to the conclusion part of each rule. This results in one fuzzy subset to be assigned to each output variable for each rule. There are two inference methods/inference rules: **MIN** and **PRODUCT**.

- **Defuzzification- Conv**ert the fuzzy value obtained from composition into a "crisp" value.

This process is often intricate since the fuzzy set might not interpret directly into a crisp value. But it considered being obligatory, since controllers of substantial systems require discrete signals. The conversion of a fuzzy quantity to a precise quantity, just as fuzzification is the conversion of a precise quantity to a fuzzy quantity. The output of a fuzzy process can be the logical union of two or more fuzzy membership functions defined on the universe of discourse of the output variable.
Defuzzification approach is intended at producing a non-fuzzy control action. The crisp value of the output variable is computed by finding the variable value of the center of gravity of the membership function for the fuzzy value. There different defuzzifying methods that used most commonly, one of them are Centroid of area (COA).

Centroid of Area (COA)

COA finds the point where a vertical line would slice the aggregate set into two equal masses. Centroid defuzzification method finds a point representing the centre of gravity of the fuzzy set on its interval.

2.2.1 Applications of Fuzzy System

Fuzzy systems nowadays are extensively worn in various domains such as Aerospace, Business, Chemical industry, Defense, Signal processing and telecommunication, Transportations, etc.

One among them is the **medical province**. Medical diagnostic support system, control of arterial pressure during anesthesia, multivariable control of anesthesia, fuzzy inference diagnosis of diabetes are some of the paradigm of fuzzy system in the medical world.

3. Optimization

The analytical methods that are used to stumble on the optimum solution or unimpeded maxima or minima of constant and differentiable function are said to be the Classical optimization techniques [7]. The formula behind these techniques is executed iteratively by comparing diverse solutions in order to acquire the expected optimal result. There are two discrete types of optimization algorithms generally used.

- **Deterministic Algorithms** - Specific rules for moving one solution to other.
- **Stochastic Algorithms** - Probabilistic translation rules for gaining popularity due to certain properties. One of such algorithm is the Swarm Intelligent (SI).

Swarm Intelligent Optimization

The synthetic intelligence which is based on the collective performance of decentralized and self-organized systems is known as Swarm Intelligent (SI). The SI is a loosely structured collection of interacting agents which can be distinguished, communicated and/or interrelated with each other. Since the agents can be easily added or removed without influencing the composition of the system, it is measured to be flexible and can be adapted in new situations [10].

Particle Swarm Optimization (PSO)

A Swarm Intelligent technique that searches for a best possible solution in the computable search space based on a population. This stochastic optimal search has been inspired by the Swarms of Bees, Flocks of Birds and Schools of Fish. The individuals of the search attempts to improve themselves by observing and imitating their neighbors. PSO are exceedingly used to find approximate solutions to extremely complex or unfeasible numeric maximization and minimization problems [6]. The algorithm of this optimization technique works as follows:

- Initialization by assembling of the random solutions.
- Searches for optimal by updating generations.
- Particles are swarmed in the solution space and the evaluating each solution are done with respect to some fitness measures behind each time step.
- In every iteration, each particle is rationalized as

✓ The first one is said to be the best fitness that has been acquired so far by the algorithm. This value is called **pbest**.
✓ Another "best" value that has been obtained by swarming among the population. This second best value is a global best and called **gbest**.
✓ When a particle takes part of the population as its topological neighbors, the second best value is a local best and is called **lbest**.

Application of PSO

PSO relevance their therapeutic work in Human tremor analysis and other deadly diseases. Human performance assessment and Ingredient mix optimization are claimed through this modus operandi.

4. Application Collaboratting the Above Techniques

According to aetiology, Atherosclerosis is alleged to be the most assassinating disease in the majority of developed and developing countries like India. Atherosclerosis is a medical terminology used to describe the ruptures of the arteries of the heart muscles by causing blood clots or plaque. The plaque is made up of fat, cholesterol, calcium and other substances

which build upon the walls of the blood vessels. The arteries are responsible for supplying the oxygen rich blood to the heart muscles. When the blood clot grows huge enough, it blocks the flow of oxygen affluent blood to the heart muscles absolutely. This causes angina or heart attacks or even to sudden death. The angina is the chest pain or uneasiness due to the lack of oxygen in the blood.

The **Coronary Heart Disease (CHD)** is the damage in the interior of the coronary arteries leading to heart attacks and Arrhythmias. The Arrhythmias is the problem even in adolescent people with respective to the rate or rhythm of their heartbeat. The early prediction of such tedious disease can reduce the mortality rate. These types of prophecy nowadays are quite impressive in the world of robotic technology. The following inspection provides us with better way to end with such prediction proposals [1].

5. Exploration of Techniques Used

S.No	*Author of the paper*	*Methodology used*	*No. of attribute*	*Attained accuracy % ~*
1	Persi Pamela et al.	CART decision Tree; Fuzzy System ;PSO	14	94%
2	Markos etal.	Decision Tree; Fuzzy modeling & Optimization	19	73.4%
3	Kantesh et al.	Fuzzy reasoning	6	80%
4	K Cinetha et al.	Fuzzy logic; Decision Tree with Clustering	1230*	97.67%
5	Debabrata et al.	CAD Screening Expert System; Fuzzy System	7	84.20%
6	S Muthukarupp-an et al.	Decision Tree; PSO; Fuzzy expert System	13	93.27%
7	Dursan et al.	Support Vector Machine ; Decision Tree; Neural Networks	19	87.74%
8	Ilias et al.	Support Vector Machine	198#	77%
9	Chih-Lin Chi et al.	Decision Support System with Optimal Decision path finder	49	50%
10	Rajeshwar et al.	Artificial Neural Network ; Back Propagation	12	89.4%

*Training data #Heart sound signal

Persi Pamela *et al*. used Classification and Regression Tree (CART) algorithm along with the Particle Swarm Optimization (PSO) to predict 94% accuracy of CHD [18]. Markos *et al.* has used C4.5 decision tree algorithm besides fuzzy optimization techniques to acquire the result of 73.4 % [15]. 1230 training data are facilitated by *K* Cinetha *et al.* to propose a Decision Support System (DSS) for precluding CHD using decision tree with Clustering techniques has attained a premier accuracy of 97.67% [5]. To envisage 87.74 % accuracy by means of 19 attributes Dursan *et al.* has used C5 and CART decision trees [8]. Fuzzy expert system along with decision tree algorithm has been intended to gain the accuracy of 93.27 % by S Muthu karuppan *et al.* [17].

6. Conclusions

From this brief study, the prediction system which incorporates the methodologies such as Data mining, Fuzzy logic and Decision tree with clustering provides us with the appropriate accuracy of 97.67%. The another system that have been implemented in the Mat lab 10 with the techniques of Decision tree , Fuzzy system and Particle Swarm Optimization provides us the best accuracy of 94% even with less number of attributes which accounts to 14. Thereby, the study can be concluded as the system proposed by Persi Pamela *et al.* as worthy in regards of performance metrics as number of attributes, accuracy rate and time consumption.

References

[1] Ahmet Yardimci, (2009), *"Soft computing in medicine"*, Applied soft Computing, Pp: 1029-1043.
[2] Anooj, P. K., (2012), *"Clinical decision support system: Risk level prediction of heart disease using weighted fuzzy rules"*, Journal of King Saud University-Computer and Information Sciences, Pp: 27-40.
[3] Breiman, L., Friedman, J.H., Olshen, R.A., and Stone, C.I. (1984), Classification and regression trees. elmont, Calif.: Wadsworth.
[4] Brijain R Patel, Kaushik K Rana, (2014), *"Use of Renyi Entropy Calculation Method for ID3 Algorithm for Decision tree Generation in Data Mining"*,International Journal of Advance Research in Computer Science and Management Studies Volume 2, Issue 5, Pp.30-34
[5] Cinetha, K. and Dr.Uma Maheswari, P., (2014), "*Decision Support System for Precluding Coronary Heart Diseases (CHD) Using Fuzzy Logic"*, International Journal of Computer Science Trends and Technology (IJCST), Pp:102-107.
[6] Debabrata Pal, K.M. Mandana, Sarbajit Pal, Debranjan Sarkar, Chandan Chakraborty, (2012), *"Fuzzy expert system approach for coronary artery disease screening using clinical parameters"*, Knowledge-Based Systems.
[7] Dong-ping Tian , Nai-qian Li, (2009), "*Fuzzy Particle Swarm Optimization Algorithm*", Proceedings of the 2009 International Joint Conference on Artificial Intelligence, Pp.263-267.
[8] Dursun Delen, Asil Oztekin, Leman Tomak, (2012), *"An analytic approach to better understanding and management of coronary surgeries"*, Decision Support Systems, Pp: 698-705.
[9] Heart disease and stroke statistics, *"Heart disease and stroke statistics update"*, American heart association, available at *http://www.americanheart.org*.
[10] Hassan M. Elragal, (2010), *"Using swarm intelligence for improving accuracy of fuzzy classifiers"*, International Journal of Electrical and Computer Engineering.
[11] Ilias Maglogiannis, Euripidis Loukis, Elias Zafiropoulos,

Antonis Stasis, (2009), *"Support vectors machine-based identification of heart valve diseases using heart sounds"*, Computer methods and programs in biomedicine, Pp: 47-61.

[12] Imran Kurt, Mevlet Ture, A. Turhan Kuram, (2008), "*Comparing performances of logistic regression, classification and regression tree and neural networks for predicting coronary artery disease"*, Expert Systems with Applications, Pp: 366-374.

[13] Jesmin Nahar, Tasadduq Imam, Kevin S. Tickle, Yi-Ping Phoebe Chen, (2013), *"Association rule mining to detect factors which contribute to heart disease in males and females"*, Expert Systems with Applications, Pp: 1086–1093.

[14] Krzysztof J. Cios, G. William Moore, (2002), *"Uniqueness of medical data mining"*, Artificial Intelligence in Medicine, Pp: 1–24.

[15] Markos G., Tsipouras *et al*., (2008), *"Automated Diagnosis of Coronary Heart Disease Based on Data Mining and Fuzzy Modeling",* Global journal of Computer Science and Technology: C Software & Data Engineering, Pp: 447 – 457.

[16] Matjaz Kukar, Igor Kononenko, Ciril Groselj, (2011), *"Modern parameterization and explanation techniques in diagnostic decision support system. A case study in diagnostics of coronary artery disease"*, Artificial Intelligence in Medicine, Pp: 77-90.

[17] Muthukaruppan, S., M.J. Er , (2012), "*A hybrid particle swarm optimization based fuzzy expert system for the diagnosis of coronary artery disease"*, Expert Systems with Applications, Pp: 11657–11665.

[18] Persi Pamela. I, Gayathri.P and N. Jaisanker , (2013), "*A Fuzzy Optimization Technique for the Prediction of Coronary Heart Disease Using Decision Tree",* International Journal of Engineering and Technology (IJET), Pp: 2506-2514.

[19] Rajeswari, K., Dr. Vaithiyanathan,V., Dr. Neelakantan, T. R. , (2012), *"Feature selection in Ischemic heart disease identification using feed forward neural networks"*, Procedia Engineering, Pp: 1818-1823.

[20] Vahid Khatibi, Gholam Ali Montazer, (2010), *"A fuzzy-evidential hybrid inference engine for coronary heart disease risk assessment"*, Expert Systems with Applications, Pp: 8536-8542.

Authors' Profile

Sowmya N. is doing her Master's Degree Computer Science (M.Sc), Avinashilingam Institute for Home Science and Higher Education for Women University, Coimbatore. She had completed her Bachelor's Degree in Computer Science in 2013. Her areas of interests are Soft Computing and Data mining.

Dr. Mrs. R.Vijayabhanu is an Assistant Professor in the Department of Computer Science, Avinashilingam Institute for Home Science and Higher Education for Women University, Coimbatore. She has completed MCA, M.Phil and Ph.D in Computer Science. Her area of interest is Soft Computing. She has published 8 papers in International Journals and presented four papers at International conferences.