A Study on various approaches of attribute selection and heart disease prediction in medical dataset classification

Deep Kumar

Software Engineer, Igniva Solutions Private Limited, Mohali, Punjab, India

Deepkumar1343@gmail.com

Abstract:

Data mining is the process of extraction of valuable information from the dataset using different attributes. In the process of data mining different relationships between attributes have been computed that can be used for various processes of data mining that are classification, clustering and association. Classification is a branch of data mining that can be used for prediction of various classes on the basis of data mining approaches. In the process of classification various rules have been used for extraction of relationships between different attributes available in the dataset. In recent research various approaches have been proposed for classification. Medical dataset classification is an emerging area of research for prediction of various diseases to an individual based on different attributes.

Keywords: Data Mining, Classification, Heart Disease Prediction, SVM, Decision Table, Decision Tree.

1. INTRODUCTION

1.1 Data Mining: It is the process of fetching hidden knowledge from a wide store of raw data. The knowledge must be new, and one must be able to use it. Data mining has been defined as "It is the science of fetching important information from wide databases". Data Mining is used to discover knowledge out of data and present the data in an easy and understood able form. It is a process to examine large amounts of data routinely collected. It is a cooperative effort of humans and computers. Best results are achieved by balancing the knowledge of human experts in describing problems and goals with the search capabilities of computers. Two goals of data mining are prediction and description. Prediction tells us about the unknown value of future variables. On the other hand Description focuses on finding patterns describing the data that can be interpreted by Humans.

1.2 Heart Disease Prediction: The heart can be affected by diverse types of diseases most of them are dangerous for human's lives. Coronary heart diseases, cardiomyopathy and cardiovascular diseases are some examples of heart diseases.The most common type of these diseases is Coronary Arteries Disease (CAD) wherein coronary arteries hard and tight. The term Cardiovascular Disease (CVD) denotes a wide range of conditions that affect the heart and the blood vessels, and the manner in which the blood is pumped and circulated throughout the body. CVD results in severe illness, disability, and is most likely to cause death. Narrowing of the coronary arteries causes the reduction of oxygen supplied to the heart and leads to the so-called Coronary Heart Disease (CHD). A sudden blockage of a coronary artery is generally due to a blood clot which may cause a heart attack. Chest pains arise when the blood received by the heart muscles is inadequate and unconnected. There are many remarks and symptoms used by the physicians to diagnose heart diseases. Age, sex, chest pain type, blood pressure, cholesterol, fasting blood sugar, maximum heart rate, and hereditary are meaningful symptoms. Besides, other habits might be used including stress, overweight, smoking, alcohols intake and less exercise.

1.3 Signs and symptoms: Angina (chest pain) that occurs regularly with activity, after heavy meals, or at other predictable times is termed stable angina and is associated with high grade **narrowing's** of the heart arteries. The symptoms of angina are often treated with beta-blocker therapy such as **metoprolol** or **atenolol**. Nitrate preparations such as **nitroglycerin**, which come in short-acting and long-acting forms are also effective in relieving symptoms but are not known to reduce the chances of future heart attacks. Many other more effective treatments, especially of the underlying **athermanous** disease, have been developed. Angina that changes in intensity, character or frequency is termed unstable. Unstable angina may precede **myocardial infarction**. About 80% of chest pains have nothing to do with the heart.

1.4Classification In Data Mining

Classification is one kind of predictive modeling. More specifically, classification is the process of assigning new objects to predefined categories or classes. Given a set of labeled records, we build a model such as a decision tree, and predict labels for future unlabeled records. Model building in the classification process is a supervised learning problem. Training examples are described in terms of (1) attributes, which can be categorical—i.e., unordered symbolic values or numeric; and (2) class label, which is also called the predicted or output attribute. If the latter is categorical, then we have a classification problem. If the latter is numeric, then we have a regression problem. The training examples are processed using some machine learning algorithm to build a decision function such as a decision tree to predict labels of new data.

Data Mining techniques such as classification, association and clustering are generally used to extract the hidden, previously unseen knowledge from voluminous of databases. Of the various data analysis techniques, classification is a supervised machine learning technique which makes predictions about the future class instances by mapping instances of testing data to the predefined class labels which is learnt from the supplied instances of classes with class labels. There are several models in classifications such as probabilistic model, evolutionary algorithmic model etc.

2. REVIEW OF LITERATURE

Shafique, U. et al [1] this paper proposed a novel approach for data mining in healthcare. Healthcare dataset contain huge amount data that contain raw information. In the process of data mining various relation and hidden patterns have to be extracted from dataset using data mining approach. In this paper various data mining approaches have been used for data mining process using machine learning algorithms. Decision tree based, neural based and Bayesian classifier based classifier has been used in this paper for data mining of heart disease dataset prediction. Attribute selection approach using best search first has been used for data mining process that minimize number of irrelevant attribute in data mining approach. performance of these different approaches has been measured on the basis of different performance evaluation parameters.

Deepali Chandna [2] proposed a novel approach for prediction of heart disease on the basis of different parameters of prediction. In medical terminology various efforts and tests have to be done for prediction of various diseases. In this paper a novel approach has been proposed that can be used for prediction of heart disease on the basis of different attributes rather than various tests have to be made. Heart disease is a major factor for death causes in today's life. In this proposed paper a novel approach has been developed that cause to quick and efficient decision making for heart disease prediction on the basis of different attributes. Thuraisingham, B.[3]this paper stated that data mining is the process of posing queries and fetching patterns from large quantities of data using pattern matching or some other reasoning techniques. Data mining has many applications in security including for national security as well as for cyber security. Threats include in national security attacking buildings, destroying critical infrastructures such as power grids and telecommunication systems. Data mining techniques are being investigated to find out who the suspicious people are and who is capable of carrying out terrorist activities. Cyber security is involved with protecting the computer and network systems against corruption due to Trojan horses, worms and viruses. Data mining is also being applied to provide solutions such as intrusion detection and auditing. Thuraisingham, B.et al [4] want to propose that the presentation will provide an overview of datamining and security threats and then discuss the applications of data mining for cyber security and national security including in intrusion detection and biometrics. Privacy considerations including a discussion of privacy preserving datamining will also be given.Asghar, S.et al [5]this paper proposed that data mining has emerged as one of the major research domain in the recent decades in order to extract implicit and useful knowledge. This knowledge can be comprehended by humans easily. This knowledge extraction was computed and evaluated manually using statistical techniques. Subsequently, semi-automated data mining techniques emerged because of the advancement in the technology. Such advancement was also in the form of storage which increases the demands of analysis. In such case, semi-automated techniques have become inefficient. So automated data mining techniques were introduced to synthesis knowledge efficiently.

3. APPROACHES USED

Decision Table Classifier: It describes two variants of decision table classifiers based conceptually on a simple lookup table. The rest classier, called DTM aj (Decision Table Majority) returns the majority of the training set if the decision table cell matching the new instance is empty, i.e., it does not contain any training instances. The second classier, called DTL oc(Decision Table Local), is a new variant that searches for a decision table entry with fewer matching attributes(larger cells) if the matching cell is empty. This variant therefore returns an answer from the local neighbor-hood, which we hypothesized, will generalize better for real datasets that tend to be smooth, i.e., small changes in a relevant attribute do not result in changes to the label.

Decision Trees: Decision trees are the best-known classification paradigm. A decision tree represents a set of classification rules in a tree form. Each root–leaf path corresponds to a rule of form $T_{i1} \wedge \dots \wedge T_{il} \rightarrow (C = c)$, where c is the class value in the leaf and each T_{ij} is a Boolean-valued test on attribute A_{ij}. The earliest decision trees were constructed by human experts, but nowadays they are usually learnt from data. The best known algorithms are ID3 and C4.5. The basic idea in all learning algorithms is to partition the attribute space until some termination criterion is reached in each leaf. Usually, the criterion is that all points in the leaf belong to one class. However, if the data contains inconsistencies, this is not possible. As a solution, the most common class among the data points in the leaf is selected. An alternative is to report the class probabilities according to relative frequencies in the node.

Bayesian Classifiers: In Bayesian networks, statistical dependencies are represented visually as a graph structure. The idea is that we take into account all information about conditional independencies and represent a minimal dependency structure of attributes. Each vertex in the graph corresponds to an attribute and the incoming edges define the set of attributes, on which it depends. The strength of dependencies is defined by conditional probabilities. For example, if A1 depends on attributes A2 and A3, the model has to define conditional probabilities P (A1|A2, A3) for all value combinations of A1, A2 and A3. When the Bayesian network is used for classification, we should first learn the dependency structure between explanatory attributes A1, AK and the class attribute C. In the educational technology, it has been quite common to define an ad hoc graph structure by experts. However, there is a high risk that the resulting network imposes irrelevant dependencies while skipping actually strong dependencies. When the structure has been selected, the parameters are learnt from the data.

K-Nearest Neighbor Classifiers: K-nearest neighbor classifiers represent a totally different approach to classification. They do not build any explicit global model, but approximate it only locally and implicitly. The main idea is to classify a new object by examining the class values of the K most similar data points. The selected class can be either the most common class among the neighbors or a class distribution in the neighborhood. The only learning task in K-nearest neighbor classifiers is to select two important parameters: the number of neighbors K and distance metric d. An appropriate K value can be selected by trying different values and validating the results in a separate test set. When data sets are small, a good strategy is to use leave–one–out cross-validation. If K is fixed, then the size of the neighborhood varies. In sparse areas, the nearest neighbors are more remote than in dense areas. However, defining different Ks for different areas is even more difficult. If K is very small, then the neighborhood is also small and the classification is based on just a few data points. As a result the classifier is unstable, because these few neighbors can vary a lot. On the other hand, if K is very large, then the most likely class in the neighborhood can deviate much from the real class.

Support Vector Machines: Support vector machines (SVMs) are an ideal method, when the class boundaries are non-linear but there is too little data to learn complex non-linear models. The underlying idea is that when the data is mapped to a higher dimension, the classes become linearly separable. The main advantage of SVMs is that they find always the global optimum, because there are no local optima in maximizing the margin. Another benefit is that the accuracy does not depend on the dimensionality of data and the system is very robust to over fitting. This is an important advantage, when the class boundary is non-linear. Most other classification paradigms produce too complex models for non-linear boundaries.

Linear Regression: Linear regression is actually not a classification method, but it works well, when all attributes are numeric. For example, passing a course depends on the student's points, and the points can be predicted by linear regression. In linear regression, it is assumed that the target attribute (e.g. total points) is a linear function of other mutually independent attributes. However, the model is very flexible and can work well, even if the actual dependency is only approximately linear or the other attributes are weakly correlated. The reason is that linear regression produces very simple models, which are not as risky for over fitting as more complex models. However, the data should not contain large gaps (empty areas) and the number of outliers should be small.

4. CONCLUSION

Data mining is the process of extraction of important features from the dataset so that relevant information can be extracted. In the process of data mining dataset has been used for feature evaluation that can be used for extraction of different information from the dataset. Data mining has been done for visualization, clustering and classification so that appropriate decision can be taken for future reference. In this research work heart disease prediction has been done using tree based classifier that develops pruning tree for dataset classification. In the purposed work genetic algorithm has been hybridize with J48 classifier for achieving best classification of the dataset. Various classification parameters have been analyzed in purposed work. By analyzing these performances evaluation parameter we can state that purposed approach provides better classification than simple J48 classifier.

REFRENCES

[1] Shafique, U.Majeed, F. Qaiser, H.et al. "Data Mining in Healthcare for Heart diseases", International Journal of Innovation and Applied Studies (ISSN), Vol. 10(4), pp. 1312-1322, 2015.

[2] Chandna, D. “Diagnosis of Heart Disease Using Data Mining Algorithm”, International Journal of Computer Science and Information Technologies, Vol. 5(2), pp. 1678-1680, 2014.

[3] Thuraisingham, B. "Data Mining for Malicious Code Detection and Security Applications", *International Joint Conference on Web Intelligence and Intelligent Agent Technology*,6-7,IEEE,*2009.*

[4] Thuraisingham, B, et al, “Data Mining for Security Applications”,International Conference on Embedded and Ubiquitous Computing,4 – 5,IEEE,2011.

[5] Asghar, S, et al, “Automated Data Mining Techniques: A Critical Literature Review”, *International Conference on Information Management and Engineering*, 978-0-7695-3595-1, 75 – 79, IEEE, 2009.

[6] Akhiljabbar, M, et al, “Heart Disease Prediction System using Associative Classification and Genetic Algorithm”, IEEE, 2012.

[7] Mahmood, A, et al, "Association Rules Mining Based Clinical Observation", Institute for Integrated and

Intelligent Systems (IIIS), 2014.

[8] El-DeenAhmeda, R. "Performance study of classification algorithms for consumer online shopping attitudes and behavior using data mining", Fifth International Conference on Communication Systems and Network Technologies, pp. 1344-1349,2015.

[9] PareshTanna "Using Apriori with WEKA for Frequent Pattern Mining", International Journal of Engineering Trends and Technology (IJETT), pp. 127-131, 2014.

[10] IlaPadhi"Predicting Missing Items in Shopping Cart using Associative Classification mining",International Journal of Computer Applications, pp. 8-11,2012.

[11] Ling Liu "Improving Online Shopping Experience using Data Mining and Statistical Techniques", Journal of Convergence Information Technology (JCIT), pp. 4-12, 2013.

[12] Bavisia, S. "A Comparative Study of Different Data Mining Algorithms", International Journal of Current Engineering and Technology, pp. 3248-3252, 2015.

[13] Anand, M, et al, "Customer Relationship Management using Adaptive Resonance Theory", International Journal of Computer Applications, pp. 43-47, 2013.

[14] Saranya,M, et al, "Decision Support System for CRM in Online Shopping System", International Journal of Advances in Computer Science and Technology, Vol. 3(2), pp. 148-151, 2014.

[15] Kamal, R. "Adaptive Pointing Theory (APT) Artificial Neural Network", International Journal of Computer and Communication Engineering, pp. 212-215, 2014.