A SURVEY ON PREDICTION OF HEART DISEASE PRESENCE USING DATA MINING AND MACHINE LEARNING TECHNIQUE

S. Sathya

Research Scholar, School Of Computer Science, Engineering & Applications, Bharathidasan University, Thiruchirappalli.620023.

Abstract—*Heart disease is the leading cause of death for both men and women. This is the case in the India and worldwide. More than half of all people who die due to heart disease are men. Earlier detection following the treatment would reduce the serious cause. We have lot of data with prescriptions for the patient. The help of the technology in Data mining (DM) and Machine Learning (ML) would substantially improve the diagnosis accuracy. Many researchers, in recent times, have been using several data mining or machine learning classification techniques to help the health care industry and the professionals to predict heart related diseases. This survey paper describes a focused survey of machine learning (ML) and data mining (DM) classification methods for heart disease prediction. In addition we reviewed data mining techniques and Machine learning classification algorithms, its processes, tools, related works and their different types of techniques can effectively decide whether the patient is suffering from heart disease or not. Moreover, we reviewed Evaluation Measures for different types of classification algorithms on Data Mining and Machine Learning domain. Therefore, we can easily measure the two or more algorithm with same dataset for same problem. Finally comes with best suitable algorithm for predict heart disease.*

Keywords—Heart disease, Data Mining, Machine learning, Classification, Evaluation measure.

1. INTRODUCTION

Heart diseases are the major cause of mortality globally, as well as in India. The term Heart disease includes many diseases that are diverse and specifically affect the heart and the arteries of a human being. They are caused by disorders of the heart and include the type of heart disease and heart failure. Heart diseases have emerged as the number one killer in world. It is the leading cause of death among all human being

According to the World Health Organization, every year more than 12 million deaths are occurring worldwide due to the various types of heart diseases, which is also known as term cardiovascular disease. About 25 per cent of human deaths in occur because of heart diseases. Heart diseases are affecting even young aged people around their 20-30 years of lifespan.

In India also the leading cause of death in all regions, though the numbers vary. The amount of deaths caused by heart disease is the highest in south India (25 per cent).

The Common modifiable possibility of heart disease namely physical inactivity, unhealthy diet, harmful effects of tobacco and alcohol and other habit-forming substances has been identified. Controlling the common modifiable risk factors shall help in prevention and control of cardiovascular diseases .The Healthcare industry collects huge amounts of healthcare data that contain all prescription and result in the form of Electrical Recorded Data which described about patient status such as positive or negative of Heart disease.

In present era, technology plays key role to improve human activity. In growth of Artificial Intelligence, computer are plays big game. Lot of technology appeared. Especially Data Mining and Machine Learning. Therefore, we can use those technologies to predict the heart disease earlier and accurately. Earlier detection followed by the treatment would reduce the serious cause. Therefore, we can use Data Mining/Machine Learning Classification algorithm to solve the problem.

This survey paper mainly describes about Different between machine learning (ML) and data mining (DM) supervised methods for classification problems and uses. In addition, we reviewed data mining techniques and Machine learning classification algorithms, its processes, tools, and Performance of Evaluate Measure for classification algorithm. Because of that, we can identify the best model or algorithm to decide whether the patient is suffering from heart disease or not.

Therefore, we can easily measure the two or more algorithm with same dataset for same classification problem with high accuracy. Then finally come up with best suitable algorithm for predict heart disease.

2.DATA MINING AND MACHINE LEARNING OVERVIEW

There is a lot of confusion about the terms ML, DM, and Knowledge Discovery in Databases (KDD). KDD is a full process that deals with extracting the unknown information from data.

DM is a particular step in this process—the application of specific algorithms for extracting patterns from data. The extra steps in the KDD process (data preparation, data

selection, data cleaning, incorporation of appropriate prior knowledge, and proper interpretation of the results of DM) guarantee that useful knowledge is extracted from available data.

There is a important go beyond between ML and DM. These two terms are commonly confused because they often employ the same methods and therefore overlap significantly.

Generally, ML defined as a "field of study that gives computers the ability to learn without being explicitly programmed." In simple word "Learn from Past experience". ML focuses on classification and prediction, based on known properties previously learned from the training data. ML algorithms need a goal (problem formulation) from the domain (e.g., dependent variable to predict). DM focuses on the discovery of previously unknown properties in the data. It does not need a specific goal from the domain, but instead focuses on finding new and interesting knowledge.

One can view ML as the older sibling of DM. The term data mining was introduced in late 1980s (the first KDD conference took place in 1989), whereas the term machine learning has been in use since the 1960s. Presently, the younger sibling (i.e., use of the term DM) is more popular than the older one, which might be the reason why some researchers actually label their work as DM rather than ML.

2.1 Steps Involved in KDD Process

Data Cleaning: Data cleaning is defined as removal of noisy and irrelevant data from collection.

- Cleaning in case of ***Missing values***.
- Cleaning ***noisy*** data, where noise is a random or variance error.
- Cleaning with ***Data difference recognition*** and ***Data transformation tools***.

Data Integration: Data integration is defined as heterogeneous data from multiple sources combined in a common source(Data Warehouse).

- Data integration using Data Migration tools.
- Data integration using Data Synchronization tools.
- Data integration using ETL (Extract-Load-**Transformation) process.**

Data Selection: Data selection is defined as the process where data relevant to the analysis is decided and retrieved from the data collection.

- Data selection using Neural network.
- Data selection using Decision Trees.
- Data selection using Naive bayes. Data selection using Clustering, Regression, etc.

Data Transformation: Data Transformation is defined as the process of transforming data into appropriate form required by mining procedure.

Data Transformation is a two step process:

Data Mapping: Transmission elements from source base to destination to capture transformations.

Code generation: Creation of the actual transformation program.

Pattern Evaluation: Pattern Evaluation is defined as identifying strictly increasing patterns representing knowledge based on given measures. Find ***interestingness score*** of each pattern. Uses ***summarization*** and ***Visualization*** to make data understandable by user.

Knowledge representation: Knowledge demonstration is defined as technique which utilizes visualization tools to represent data mining results.

- Generate ***reports***.
- Generate ***tables***.
- Generate ***discriminantrules***, ***classification rules***, ***characterization rules***, etc.

An ML approach usually consists of two phases: training and testing. Often, the following steps are performed:

Identify class attributes (features) and classes from training data.

Identify a subset of the attributes required for classification (i.e., dimensionality reduction).

Learn the model using training data.

Use the trained model to categorize the unknown data.

In the case of misuse detection, in the training phase each misuse class is learned by using appropriate exemplars from the training set. In the testing phase, new data are run through the model and the exemplar is classified as to whether it belongs to one of the misuse classes. If the exemplar does not belong to any of the misuse classes, it is classified as normal.

3.DATASET DESCRIPTION

The Cleveland heart dataset from the UCI Machine Learning Repository as it is generally used by the Pattern design community. The dataset consists of 303 human being clinical reports in which 164 did not have any disease. In this dataset there is a total of 97 female patients in which 25 people are the confirmatory case, also there are 206 male patients in which 114 are diagnosed with the disease.

1. age: age in years
2. sex: sex (1 = male; 0 = female)

3. cp: chest pain type

   Value 1: typical angina

   Value 2: atypical angina

   Value 3: non-angina pain

   Value 4: asymptomatic

4. trestbps: resting blood pressure (in mm Hg on admission to the hospital)
5. chol: serum cholesterol in mg/dl
6. fbs: (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)
7. restecg: resting electrocardiographic results
   - Value 0: normal
   - Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV)
   - Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria
8. thalach: maximum heart rate achieved
9. exang: exercise induced angina (1 = yes; 0 = no)
10. old peak = ST depression induced by exercise relative to rest
11. slope: the slope of the peak exercise ST segment
    - Value 1: upsloping
    - Value 2: flat
    - Value 3: downsloping
12. ca: number of major vessels (0-3) colored by fluoroscopy
13. thal: 3 = normal; 6 = fixed defect; 7 = reversable defect
14. num: diagnosis of heart disease (angiographic disease status)
    - Value 0: < 50% diameter narrowing
    - Value 1: > 50% diameter narrowing

4. RESEARCH METHODOLOGY

In this section we described different types of learning and their algorithms for both Data mining and machine learning domains.

4.1 Supervised Learning

Supervised learning is a learning in which we educate or train the machine using data which is well labeled that means some data is already tagged with the correct answer. After that, the machine is provided with a new set of examples (data) so that supervised learning algorithm analyses the training data(set of training examples) and produces a right result from labeled data.

Supervised learning classified into two categories of algorithms:

- **Classification**: A classification problem is when the amount produced variable is a category, such as "disease" or "no disease".

- **Regression**: A regression problem is when the output variable is a genuine value, such as "dollars" or "weight"

This is list of common supervised algorithm Nearest Neighbor, Naive Bayes, Decision Trees, Linear Regression, Support Vector Machines (SVM), Neural Networks

4.2 Unsupervised learning

Unsupervised learning is the preparation of appliance using in order that is neither classified nor labeled and allow the algorithm to act on that information exclusive of supervision. Here the task of engine is to collection unsorted information according to similarities, patterns and differences without any prior training of data. Unlike supervised learning, no instructor is provided that means no training will be given to the machine. Therefore machine is controlled to find the hidden structure in unlabeled data by our-self

Unsupervised learning classified into two categories of algorithms:

- **Clustering**: A clustering problem is where you want to discover the innate groupings in the data, such as grouping customers by purchasing behavior.

- **Association**: An association rule knowledge problem is where you want to find out rules that give details large portions of your data, such as public that buy X also tend to buy Y

The unsupervised learning is used to List of CommoAlgorithms is k-means clustering, Association Rules

4.3 Semi-supervised Learning

In the previous two types, either there are no labels for all the surveillance in the dataset or labels are present for all the clarification. Semi-supervised learning falls in between these two. In many practical situations, the charge to label is quite high, since it require skilled human being experts to do that. So, in the deficiency of labels in the majority of the observations but present in few, semi-supervised algorithms are the best candidates for the model building. These methods use the idea that even though the group memberships of the unlabelled data are unknown, this data carries important information about the group parameters.

4.4 Reinforcement Learning

Reinforcement Learning is a category of *Machine Learning*, and thereby also a division of *Artificial Intelligence*. It allows machines and software agents to manually determine the ideal performance within a exact context, in order to

maximize its performance. Simple return feedback is required for the agent to learn its behaviour; this is known as the reinforcement signal. The Reinforcement leaning is mostly used to **list of common algorithm** Q Learning, Temporal Difference (TD) , Deep Adversarial Networks.In order to predict the heart disease we have to use classification algorithm.

4.5 Classification Algorithms

Classification is a process discovery model (functions) that describe and distinguish classes of data or concept that aims to be used to predict the class of the object which label class is unknown. Classification is part of data mining, where data mining is a term used to describe the knowledge discovery in databases. Data mining is as well a development that uses statistical techniques, mathematics, artificial intelligence, and machine learning for extracting and identifying useful information and relevant knowledge from a variety of large datasets.

The classification process is based on four components

Class

Categorical dependent variable in the form that represents the 'label' contained in the object. **For example:** heart disease risk, credit risk, customer loyalty, the type of earthquake.

Predictor

The independent variables are represented by characteristic (attribute) data. **For example:** smoking, drinking alcohol, blood pressure, savings, assets, salaries.

Training dataset

One data set that contains the value of both components above are used to determine a suitable class based on predictor.

Testing dataset

Containing new data which will be classified by the model that has been

5. EVALUTAION MEASURE FOR CLASSIFICATION ALGORITHM

In this section, we described evaluation measure for performing classification algorithm.

We have lot of classification algorithm, each algorithm has own methods to classify the data. So we need to find the best classification algorithm. That is our ultimate goal.

In order to achieve, we need some measurement that is done by using confusion Matrix.

The general view of confusion matrix is given below.

Confusion Matrix

	Predicted Class	
	Yes	No
Actual class		
Yes	TP	FN
No	FP	TN

In confusion matrix the predicted class is the class that is predicted by the classifier and the actual class is the class that is given in the data set.

- **True positives (TP):** These refer to the positive tuples that were right labeled by the classifier. Let TP be the number of true positives.
- **True negatives (TN):** These are the negative tuples that were suitably labeled by the classifier. Let TN be the number of true negatives.
- **False positives (FP):** These are the negative tuples that were wrongly labeled as positive (e.g., tuples of class buys computer = no for which the classifier predicted buys computer= yes). Let FP be the number of false positives.
- **False negatives (FN):** These are the optimistic tuples that were mislabeled as negative (e.g., tuples of division buys computer = yes for which the classifier predicted buys computer= no). Let FN be the number of false negatives

Based on confusion matrix we can determine the performance analysis. They are the list of measure with details.

Accuracy: Classification accuracy is the percentage of instances that are correctly classified by the model. It is calculated as the sum of correct classification divided by the total number of samples.

$$\text{Accuracy} = \frac{TP + FN}{TP + FP + TN + FN}$$

Sensitivity: It is the measure of the ability of a classification model to select instances of certain class from the dataset. It is the proportion of actual positive which are predicted positive.

$$\text{Sensitivity} = \frac{TP}{TP + FN}$$

Specificity: This is a measure that is commonly used in two class problems where the focus is on a particular class. It is the proportion of the negative class that was predicted negative and it is also known as the true negative rate.

$$\text{Specificity} = \frac{TN}{FP + TN}$$

Precision : Precision is the ratio of correctly predicted positive instance to the total predicted positive instance.

Precision = TP/TP+FP

$$\text{Precision} = \frac{TP}{TP + FP}$$

Recall (Sensitivity) - Recall is the ratio of correctly predicted positive instance to the all instance in actual class - yes.

$$\text{Recall} = \frac{TP}{TP + FN}$$

F measure – F measure is the weighted average of Precision and Recall. Therefore, this measure takes both false positives and false negatives into account.

$$\text{F Measure} = \frac{2 * (\text{Recall} + \text{Precision})}{(\text{Recall} + \text{Precision})}$$

Mean Absolute Error: Measure of difference between two continuous variables.

$$\text{MAE} = \frac{\sum_{i=1}^{n} Actual_i - Forecast_i}{n}$$

Root Mean Squared Error: It follows an assumption that error are unbiased and follow a normal distribution.

$$\text{RMS} = \sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n}}$$

ROC (Receive operating characteristic): It is a graphical representation of performance of classifiers

5. OPEN SOURCE TOOLS FOR DL AND ML

Weka: Weka is a anthology of machine learning algorithms for data mining tasks. The algorithms can either be practical directly to a dataset or called from your own Java code. Weka contains tools for data pre-processing, classification, degeneration, clustering, association rules, and visualization. It is also well-suited for emergent new machine learning schemes.

Orange: - Python is selection up in popularity because it's simple and easy to learn yet great. Hence, when it comes to looking for a implement for your work and you are a Python developer, look no further than Orange, a Python-based, powerful and open source tool for both novices and experts.

TensorFlow: in the beginning released in 2015, TensorFlow is an open source machine-learning framework that is easy to use and deploy across a variety of platforms. It is one of the most well maintained and extensively used frameworks for machine learning. produced by Google for sustaining its research and production objectives, TensorFlow is presented in Python, C++, Haskell, Java, Go, Rust, and most recently, JavaScript. You can also find third-party junk mail for other programming languages. The structure allows you to develop neural networks (and even other computational models) using flow graphs.

Scikit-learn: Initially released in 2007, scikit-learn is an open source library industrial for machine learning. This traditional agenda is written in Python and features several machine learning models including classification, regression, clustering, and dimensionality reduction.

Scikit-learn is designed on three other open source projects—Matplotlib, NumPy, and SciPy—and it focuses on data mining and data analysis.

Theano: at the start released in 2007, Theano is an open source Python library that allows you to easily fashion various machine-learning models. At its core, it enables you to simplify the process of defining, optimizing, and assessing mathematical expressions. Theano is capable of taking your structures and transforming them into very well-organized code that integrate with NumPy, efficient native libraries such as BLAS, and native code (C++).

Furthermore, it is optimized for GPUs, provides efficient symbolic differentiation, and comes with extensive code-testing capabilities.

Caffe: Initially at large in 2017, Caffe (Convolutional Architecture for Fast Feature Embedding) is a machine learning structure that focuses on perspicuity, speed, and modularity. The open source framework is printed in C++ and comes with a Python interface. Caffe's main features include an expressive structural design that inspires innovation, extensive code that facilitates active

development, fast presentation that accelerates industry operation, and a vibrant community that stimulates growth.

Torch: Initially at large in 2002, Torch is a machine-learning library that offers a wide array of algorithms for full of meaning learning. The open source structure provides you with optimized flexibility and speed when handling machine-learning projects—without causing avoidable complexities in the process. It is written using the scripting language Lua and comes with an original C implementation. The Torch is used to N-dimensional arrays, linear algebra routines, numeric optimization routines, well-organized GPU support, and continue for iOS and Android platforms.

Accord.NET: To begin with free in 2010, Accord.NET is a machine-learning framework entirely written in C#.

The open resource framework is suitable for production-grade scientific computing. With its extensive range of libraries, you can build various application in artificial neural networks, arithmetical data processing, image processing, and various others.

6. CONCLUSIONS

The present trend is many of the researches used many data mining and machine learning techniques' to predict the disease.

However, day-to-day many algorithms are proposed and used. Nevertheless, we could not tell which the best is.

The general objective is to study the a range of data mining techniques available to predict the heart disease and to compare them to find the best method of prediction. Supervised learning classification algorithms are usually describe as performing the task of searching through a theory space to find a suitable model that will make high-quality predictions with a exacting problem. In this paper described list of technology available in machine learning and data mining domain and their different uses. Also described the evaluation measure for classification algorithm and tools are exist in open source. Therefore, we can compare the each algorithm and find the best algorithm to predict the Heart disease with high accuracy.

7.FUTURE SCOPE

In future, we are planning to introduce an efficient disease prediction system to predict the heart disease with better accuracy utilizing different data mining and machine learning classification techniques.

8.REFERENCES

[1] A. Floares., A. Birlutiu. "Decision Tree Models for Developing Molecular Classifiers for Cancer Diagnosis". WCCI 2012 IEEE World Congress on Computational Intelligence June, 10-15, 2012 - Brisbane, Australia.

[2] Adem Karahoca, Dilek Karahoca and Mert Şanver, "Data Mining Applications in Engineering and Medicine", ISBN 978-953-51-0720-0, In Tech, August 8, 2012.

[3] B. Chaudhuri and U. Bhattacharya." Efficient training and improved performance of multilayer perceptron in pattern classification". Neuro computing, 34, pp11–27, September 2000.

[4] BalaSundar V, T Devi and N Saravan, "Development of a Data Clustering Algorithm for Predicting Heart", International Journal of Computer Applications, vol. 48, pp. 423-428,2012

[5] David L. Olson, DursunDelen, YanyanMeng, "Comparative analysis of data mining methods for bankruptcy prediction", Decision Support Systems vol.52, pp.464–473, 2012

[6] SellappanPalaniappan, RafiahAwang "Intelligent Heart Disease Prediction System Using Data Mining Techniques" IEEE, pp.978-1-4244-1968,2008

[7] SajidaPerveen, Muhammad Shahbaz, Aziz Guergachi, Karim Keshavjee, "Performance Analysis of Data Mining Classification Techniques to Predict Diabetes", Procedia Computer Science vol.82, pp.115 – 121,2015

[8] Ho, T.B. (nd). "Knowledge Discovery and Data Mining Techniques and Practice". 2006, Available on: www.netnam.vn/unescocourse/knowlegde/knowfrm.htm

[9] http://archive.ics.uci.edu/ml (The UCI Machine Learning Repository is a collection of databases).

[10]www.searo.who.int/india/cardiovascular_diseases/Commision_on_Macroeconimic