Hybrid Learning Model for Anomaly Detection and Automatic Labeling for Heart Disease Prediction.

Nivedhitha.V Assistant professor,
Department of Computer Science and Engineering, SSM Institute of Engineering and Technology, Dindigul, Tamilnadu, India.

Avanthiga.S.K , MahaLakshmi.P ,Vimalakarthika.P

Department of Computer Science and Engineering, SSM Institute of Engineering and Technology, Dindigul, Tamilnadu, India.

.

ABSTRACT

This project proposes a Hybrid Learning Model which uses both Clustering and Classification methods (HLMCC) to automate the labelling process and detect anomalies in data mining. The model consists of two practical phases, automatic labelling and detecting anomalies. First the HLM groups the data into normal labelled one and unlabelled data clusters by adopting Hierarchical Affinity Propagation (HAP) clustering. Second, the labelled data obtained from the clustering phase is used to train the Decision Trees (DTs) and to classify future unseen data. The results show that the HLM is able to automate the labelling of data, which is beneficial to minimize human involvement.

.

Keywords:. Hybrid Learning, HAP,DT,HLM.

1. INTRODUCTION

The IoT has been found in several application domains such as smart homes, wearable devices, smart cities, health care, agriculture, transportation, and industrial sectors of industry. IoT devices generate data that may behave inconsistently owing to abnormal or anomaly behavior as a result of attack issues or breakdown in devices, as examples. An anomaly, in this context, means an abnormality in the data that differs from the predicted pattern.

The characteristics of an anomaly are: different from the norm and occurring rarely in the Data Anomaly detection is the technique of identifying rare observations which do not follow the expected behavior. The major technique for performing anomaly detection involves the use of machine learning algorithms. This helps to improve the performance of the system by learning from and using data from previous experiences. There are three types of machine learning task, which are supervised, unsupervised, and semi-supervised learning. Supervised learning trains the model based on predefined labeled data, while unsupervised learning similarities between unlabelled data. Semi-supervised learning deals with partially labeled data to build the model.
Most current anomaly detection systems rely on labeled data which may not be available or it is time-consuming and expensive to produce. In addition, the data collected from IoT devices usually lack the class label and form as unlabelled data. Moreover, the volume of IoT data is growing at an increasingly rapid rate, creating a need to predict, detect, and classify any anomaly for future unseen data. To overcome these limitations, this paper proposes a Hybrid Learning anomaly detection Model that employs Clustering and Classification approaches called HLMCC.

The HLMCC model consists of two functional phases: automatic labeling and detecting anomalies. In the automatic labeling phase, Hierarchical Affinity Propagation (HAP) clustering is applied to

automate the labeling process, which helps to address the issue of unlabelled data and can be helpful in reducing human intervention. In detecting anomalies, the obtained labeled data is used to train the Decision Trees (DTs) to detect and classify future unseen data.

The main contributions of this text are as follows:

• To propose the HLMCC model based on clustering and classification approaches to automate data labeling and to detect anomalies in IoT.

• Label the data by employing the HAP clustering algorithm.

• Compare HLMCC against the DTs on the originally labeled data and the existing model.

2. LITERATURE SURVEY

To construct a machine learning model, there are diverse elements which should be considered, such as datasets, the type of learning algorithm, feature selection and evaluation techniques. For anomaly detection, the data is collected from IoT devices and placed into data storage. Then, machine learning techniques (supervised, unsupervised or semi-supervised) are implemented. Finally, validation techniques are used to evaluate the performance of the model. The existing solutions based on machine learning algorithms (supervised, unsupervised or semi-supervised) for anomaly detection in IoT are described in the following sub-sections. Data mining provides the methodology and technology to alter these mounds of data into useful information for decision making. By using data mining techniques it takes less time for the prediction of the disease with more accuracy. Among the increasing research on heart disease predicting system, it has happened to significant to categories the research outcomes and gives readers with an outline of the existing heart disease prediction techniques in each category. In this paper we study different papers in which one or more algorithms of data mining used for the prediction of heart disease. As of the study it is observed that Fuzzy Intelligent Techniques increase the accuracy of the heart disease prediction system.[1]

A. Supervised Learning

Supervised learning builds a model based on predefined labelled data. Training and testing are the two phases for supervised learning. In the training phase, we build the model using the training data, while in the testing step, the trained model provides the class label for unseen data. There is a wide range of learning algorithms such as Neural Networks (NNs), Support Vector Machines (SVMs) and K-nearest neighbors (KNNs).Different learning approaches such as single, ensemble or hybrid models are explained by Tsai that are used to classify the data into either normal or abnormal classes. Single models consist of a single classifier such as KNNs, SVMs, NNs, whereas ensemble models improve the system performance by including diverse weak classifiers, while hybrid models consist of more than two classifiers in a model such as the neuro-fuzzy model. Two main steps are performed in hybrid models are,

• The model uses the data to produce the intermediate results.

• Secondly, the intermediate results are used as input to output the final results.

This paper summarizes some of the current exploration on heart disease prediction using data mining algorithms, analyze and compare them to conclude which technique is more effective and efficient.[2]

B. Unsupervised Learning

In some pattern recognition problems, the training data consists of a group of input vectors x with none corresponding target values. The goal in such unsupervised learning problems could also be to get

groups of comparable examples within the info , where it's called clustering, or to work out how the info is distributed within the space, referred to as density estimation. To put forward in simpler terms, for a n-sampled space x1 to xn, true class labels are not provided for each sample, hence known as learning without teacher. Unsupervised Learning is harder as compared to Supervised Learning tasks..Annotating large datasets is very costly and hence we can label only a few examples manually. Example: Speech Recognition. Many researchers have proposed the use of K-nearest neighbor (KNN) algorithm for diabetes disease prediction. Some researchers have proposed a different approach by using Kmeans clustering for preprocessing and then using KNN for classification.[3]

Unsupervised Learning can be further classified into two categories:

• Parametric

• Non-Parametric

The critical factors that are mandatory for occurrence of coronary heart disease are taken at first level and the rest one are taken at second level. This two level approach increases the performance of our work as it helps in predicting disease chances accurately. The heart disease dataset is taken from UCI machine learning repository to train the neural network and then fuzzy rules are applied to predict the chances of coronary heart disease as low, medium or critical.[5]

Parametric Unsupervised Learning: It assumes that sample data comes from a population that follows a probability distribution based on a fixed set of parameters. Theoretically, in a normal family of distributions, all members have the same shape and are parameterized by mean and standard deviation. That means if you recognize the mean and variance , which the distribution is normal, you recognize the probability of any future observation. Dempster's rule of combination combines three beliefs to arrive at one final decision. Our experiments with k-fold cross validation show that the nature of the data set has a Bigger impact on some classifiers than others and the Classification based on combinedbelief shows better Overall accuracy than any individual classifier.[6] Parametric Unsupervised Learning deals withconstruction of Gaussian Mixture Models and Expectation-Maximization algorithm to find the categories in question. This case is far harder than the quality supervised learning because there are not any answer labels available and hence there's no correct measure of accuracy available to see the result.

Non-parametric Unsupervised Learning: In non-parameterized version of unsupervised learning, the data is grouped into clusters, where each cluster says something about categories and classes present in the data. In the health care industry the data mining is mainly used for predicting the diseases from the datasets.The huge amounts of data generated for prediction of heart diseases are too complex. Data mining provides methods and techniques to transform these mounds of data into useful information for decision making.[7] This method is commonly used to model and analyze data with small sample sizes. Unlike parametric models, nonparametric models don't require the modeler to form any assumptions about the distribution of the population, then are sometimes mentioned as a distribution-free method.

The prognosis of life for patients with heart failure remains poor. By using data mining methods, the purpose of this study was to evaluate the most important criteria for predicting patient survival and to profile patients to estimate their survival chances together with the most appropriate technique for health care. Five hundred and thirty three patients who had suffered from cardiac arrest were included in the analysis. We performed classical statistical analysis and data mining analysis using mainly Bayesian networks.[8]

C. Semi-supervised Learning

In this sort of learning, the algorithm is trained upon a mixture of labeled and unlabeled data. Typically, this combination will contain a very small amount of labeled data and a very large amount

of unlabeled data. The basic procedure involved is that first, the programmer will cluster similar data using an unsupervised learning algorithm then use the prevailing labeled data to label the remainder of the unlabeled data. On this data, 83.7% predictions were correct, exceeding the results obtained using the standard Support Vector Machine and equivalent kernels.[9] The typical use cases of such sort of algorithm have a standard property among them – The acquisition of unlabeled data is comparatively cheap while labeling the data is very expensive. The proposed method is an associative classifier based on the efficient FPgrowth method. Since the volume of patterns produced can be large, we offer a rule cohesion measure that allows a strong push of pruning patterns in the pattern-generating process.[10]

3. IMPLEMENTATION

The conceptual phases of the HLMCC model are shown in Figure 1. The HLMCC model consists of two phases:

1. Automatic Labeling: Employing HAP clustering to classify the data label into normal and abnormal clusters.

2. Detecting Anomalies: The labeled data obtained from the clustering is used to train DTs.

The HLMCC model applies Algorithm 1, which receives unlabelled data as input and classifies the data into normal and abnormal classes as output.

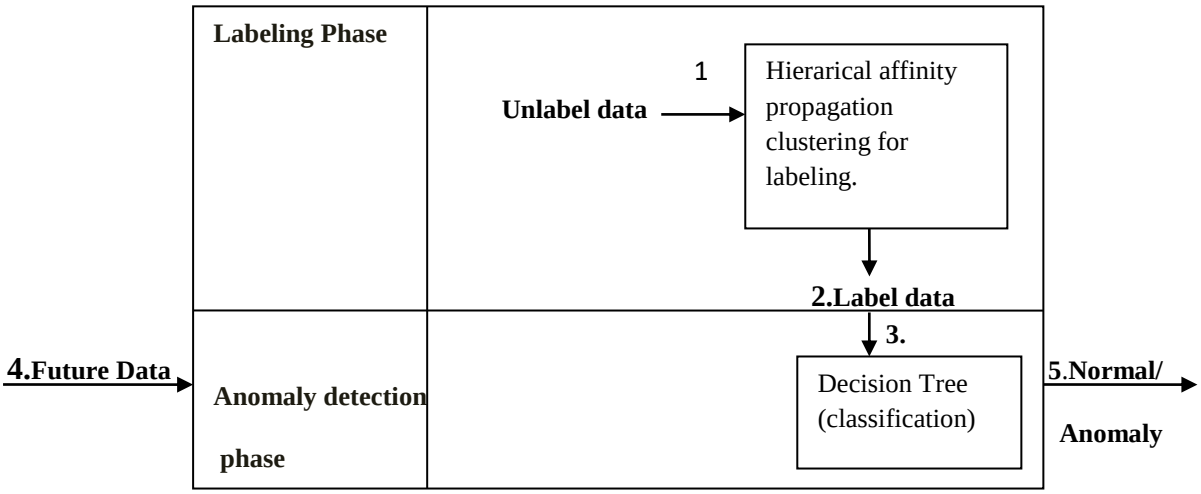

Figure 1. Conceptual phases of the HLMCC model

Algorithm 1:

Input: D: Unlabelled dataset
Output: Classified data as anomaly or normal

1: Employ HAP algorithm to cluster D into two groups,
2: Label D using the clusters, cluster 1: anomaly and cluster 2: normal,
3: Split D into two partitions Dtr for training and Dst for test,
4: Train the DTs using Dtr,
5: Classify Dst as an anomaly or a normal label using DTs.
6: End

MODULE DESCRIPTION

In this project have 8 Modules such as, Data Preprocessing, Data Cleaning, Feature Scaling, Factorization, Support Vector Machine, K -Nearest Neighbor (k-NN), Decision Tree (DT), K-Means .

Data Preprocessing: The dataset obtained is not completely accurate and error free. Hence, we'll first perform the subsequent operations there on .

Data Cleaning: NA values in the dataset are the major setback for us as it will reduce the accuracy of the prediction profoundly so, we will remove the fields which do not have values. We will substitute it with the mean of the column. This way, we will remove all the values in the data set.

Feature Scaling: Since the range of values of data varies widely, in some machine learning algorithms, objective functions won't work properly without feature scaling. For example, the bulk of classifiers calculate the space between two points by the Euclidean distance. If one among the features features a broad range of values, the space are going to be governed by this particular feature. Therefore, the range of all features should be scaled in order that each feature contributes approximately proportionately to the ultimate distance. So we'll scale the varied fields so as to urge them closer in terms of values. E.g. Age has just two values i.e. 0, 1 and cholesterol has high values like 100. So, so as to urge them closer to every other we'll got to scale them.

Factorization: In this process assigned a meaning to the values so that the algorithm doesn't confuse between them. For example, assigning aiming to 0 and 1 within the age section in order that the algorithm doesn't consider 1 as greater than 0 therein section.

Support Vector Machine: Support vector machine (SVM) is supervised learning method that analyzes data used for classification and regression analysis. It is given a group of coaching data, marked as belonging to either one among two categories; an SVM training algorithm then builds a model that assigns new examples to one category or the other, making it a no probabilistic binary linear classifier. An SVM model may be a representation of the examples as points in space, mapped in order that the samples of the separate categories are divided by a transparent gap that is as wide as possible. New examples are then mapped into that very same space and predicted to belong to a category supported which side of the gap they fall. The points are separated supported hyper plane that separate them. When data aren't labeled, supervised learning isn't possible, and an unsupervised learning approach is required, which attempts to seek out natural clustering of the info to groups, and then map new data to these formed groups. In the project, we've used this algorithm to classify the patients into groups consistent with the danger posed to them supported the parameters provided.

K -Nearest Neighbor (k-NN): K-nearest neighbor (k-NN) is one of the modest and conventional nonparametric techniques for classifying samples It calculates the approximate distances between various points on the input vectors, then assigns the unlabeled point to the category of its K-nearest neighbors. In the process of making k-NN classifier, (k) is a crucial parameter and various (k) values can cause various performances. If k is extremely huge, the neighbors, which used for prediction, will consume large classification time and affect the prediction accuracy.

Decision Tree (DT): Quinlan defined Decision Trees as "powerful and common tools for classification and prediction. A decision tree may be a tree that has three main components: nodes, arcs and leaves. Each node is labeled with a feature attribute, which is most informative among the attributes not yet considered in the path from the root. Each arc out of a node is labeled with a feature value for the node's feature, and each leaf is labeled with a category or class. A decision tree can then be used to classify a data point by starting at the root of the tree and moving through it until a leaf node is reached.

K-Means: K-means algorithm is a traditional clustering algorithm. It divides the data into k clusters, and guarantee that the data within the same cluster are similar, while the data in a various clusters have low similarities. K-means algorithm is first selected K data at random as the initial cluster center, for the rest data add it to the cluster with the highest similarity according to its distance to the cluster center, and then recalculate the cluster center of each cluster. Repeat this process until each cluster

center doesn't change. Thus data are divided into K clusters. Unfortunately, K-means clustering is sensitive to the outliers and a set of objects closer to a centroid may be empty, in which case centroids cannot be updated.

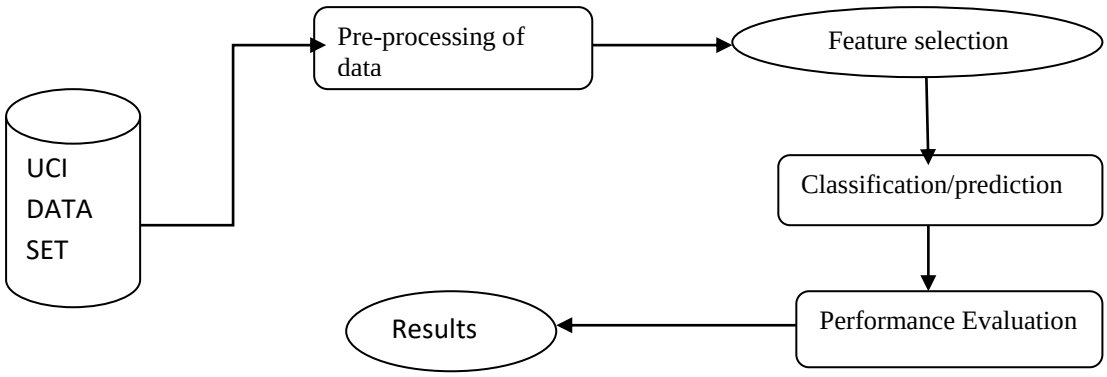

Figure 2. Module Interface Diagram

80% of the time in AI is majorly deals with preparing data.While preparing the data for learning some attacks May occur,

•A "cyber attack" refers to an attack arised through the network which will cause some economic harm.
•A "physical attack" refers to the fault arised in the physical component of the system.

1. Identify Data required.
2. Identify the availability of data, and location of them.
3. Profiling the data.
4. Source the data.
5. Integrating the data.
6. Cleanse the data.
7. Prepare the data for learning.

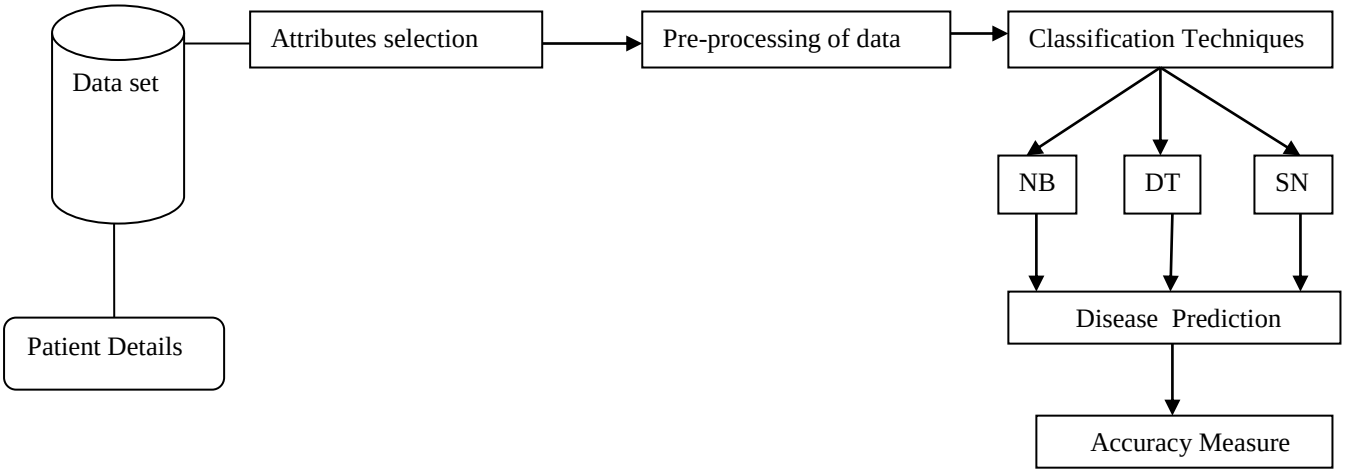

Figure 3. Over All System Architecture

The steps below are happening while performing anomaly detection.

- All the dataset about the patient details has its own property.
- Based on that attribute selection is done which will be used by the further process called pre-processing on data.

- Pre-processing deals with availability of data, cleaning the data, feature scaling factorization etc…
- Classification techniques are implemented for predicting the diseases.
- Based on multiple prediction algorithms performance should be evaluated.

DATA SET: HEART DISEASE PREDICTION TECHNIQUES

Datasets: a set of instances may be a dataset and when working with machine learning methods we typically need a couple of datasets for various purposes.

Training Dataset: A dataset given to the machine learning algorithm for training the model.

Testing Dataset: A dataset that we use to validate the accuracy of our model but not to training the model. It may be called the validation dataset.

ATTRIBUTES

ATTRIBUTES	**PROPERTIES**
age	age in years
sex	(1=male;0= female)
cp	chest pain type
trestbps	Resting blood pressure (in mm Hg on admission to the hospital)
chol	serum cholesterol in mg/dl
fbs	(fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)
restecg	resting electrocardiographic results
thalach	maximum heart rate achieved
exang	exercise induced angina(1=yes;0=no)

oldpeak	ST depression induced by exercise relative to rest
slope	the slope of the peak exercise ST segment
ca	number of major vessels(0-3) colored by fluoroscopy
thal	3 = normal; 6 = fixed defect; 7 = reversible defect
target	1or 0

4.CONCLUSION

The aim of this project is to design and develop the GSA-based AIDS as an effective IDS. The proposed system produces higher-quality labeled data by implementing ensemble clusters with an efficient clustering technique, and enhancing the capability of the detection classifier by utilizing an efficient method. The ensemble clusters technique was designed to convert unlabeled data to higher-quality labeled data and the GSA-SVM classifier was designed to enhance the classification process in the detection classifier. The results showed an improvement in detection effectiveness when using the high-quality labeled dataset which scored 97.0% on the overall accuracy and 0.03 % on the false positive rate. The GSA-based AIDS using the higher-quality labeled dataset outperformed the GSA-based AIDS using the KDD 99 test dataset. The detection accuracy improved by 10.84 % while the false positive rate reduced by 0.17 % when using a higher-quality labeled dataset. The detection accuracy of the GSA-based AIDS using the GSA-SVM classifier improved by 6.95 % ,while the false positive rate reduced by 0.07 % as compared to the GSA-based AIDS using the SVM classifier.

REFERENCES

1. V. Krishnaiah, G. Narasimha, N. Subhash Chandra, “Heart Disease Prediction System using Data Mining Techniques and Intelligent Fuzzy Approach: A Review” IJCA 2016.

2. K.Sudhakar, Dr. M. Manimekalai “Study of Heart Disease Prediction using Data Mining”, IJARCSSE 2016.

3. NagannaChetty, Kunwar Singh Vaisla, NagammaPatil, “An Improved Method for Disease Prediction using Fuzzy Approach”, ACCE 2015.

4. VikasChaurasia, Saurabh Pal, “Early Prediction of Heart disease using Data mining Techniques”, Caribbean journal of Science and Technology,2013

5. ShusakuTsumoto,” Problems with Mining Medical Data”, 0-7695- 0792-1 I00@ 2000 IEEE.

6. Y.Alp Aslandoganet. al.,” Evidence Combination in Medical Data Mining”, Proceedings of the international conference on Information Technology: Coding and Computing (ITCC’04) 0-7695-2108-8/04©2004 IEEE.

7. Carlos Ordonez, "Improving Heart Disease Prediction Using Constrained Association Rules," Seminar Presentation at University of Tokyo, 2004.

8. Franck Le Duff, CristianMunteanu, Marc Cuggiaa, Philippe Mabob, "Predicting Survival Causes After Out of Hospital Cardiac Arrest using Data Mining Method", Studies in health technology and informatics, Vol. 107, No. Pt 2, page no. 1256-1259, 2004.

9. Boleslaw Szymanski, Long Han, Mark Embrechts, Alexander Ross, KarstenSternickel,Lijuan Zhu, "Using Efficient Supanova Kernel For Heart Disease Diagnosis", Proc. ANNIE 06, intelligent engineering systems through artificial neural networks, vol. 16,page no. 305-310, 2006.

10. Kiyong Noh, HeonGyu Lee, Ho-Sun Shon, Bum Ju Lee, and Keun Ho Ryu, "Associative Classification Approach for Diagnosing Cardiovascular Disease", Springer 2006,Vol:345, page no. 721- 727.