International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Data Mining Techniques in Healthcare

[1]Anusuya Devi A., [2]Bhuvaneswari P, [3]Hansiha M, [4]Gayathri. K, [5]Loganayaki S.

[1,2,3,4,5] **Sri GVG Visalakshi College for Women**

ABSTRACT

Data Mining is one of the most motivating area of research that is become increasingly popular in health organization. Data Mining plays an important role for uncovering new trends in healthcare organization which in turn helpful for all the parties associated with this field. This survey explores the utility of various Data Mining techniques such as classification, clustering, association, regression in health domain. In this paper, we present a brief introduction of these techniques and their advantages and disadvantages. This survey also highlights applications, challenges and future issues of Data Mining in healthcare. Recommendation regarding the suitable choice of available Data Mining technique is also discussed in this paper.

INTRODUCTION

Data mining is an assortment of algorithmic techniques to extract instructive patterns from raw data. Healthcare industry today produces huge amounts of multifarious data about hospitals, resources, disease diagnosis, electronic patient records, etc. The large amount of data is crucial to be processed and scrutinized for knowledge extraction that empowers support for understanding the prevailing circumstances in healthcare industry. Data mining processes include framing a hypothesis, gathering data, performing pre-processing, estimating the model, and understanding the model and draw the conclusions [2]. Before studying how data mining algorithms are being applied on medical data, let us understand what types of algorithms exists in data mining and how they are functioning.

DATA MINING TECHNIQUES

The extraction of significant patterns from the heart disease data warehouse was presented. The heart disease data warehouse contains the screening clinical data of heart patients. Initially, the data warehouse preprocessed to make the mining process more efficient. The first stage of Association Rule used preprocessing in order to handle missing values. Later applied equal interval binning with approximate values based on medical expert advice on Pima Indian heart attack data. The significant items were calculated for all frequent patterns with the aid of the proposed approach. The frequent patterns with confidence greater than a predefined threshold were chosen and it was used in the design and development of the heart attack prediction system. Each data mining technique serves a different purpose depending on the modeling objective.

ASSOCIATION ALGORITHM

Association is a data mining technique that discovers the probability of the co-occurrence of items in a collection. The relationships between co-occurring items are expressed as Association Rules.

Association Rules for Healthy Heart:

Medical dataset:

Attribute	Description
AGE	Ageofpatient
LM	LeftMainnarrowing
LAD	LeftAnteriorDesc.arterynarrowing
LCX	LeftCircumflexarterynarrowing
RCA	RightCoronaryarterynarrowing
AL	Antero-Lateral
AS	Antero-Septal
SA	Septo-Anterior
SI	Septo-Inferior
IS	Infero-Septal
IL	Infero-Lateral
LI	Latero-Inferior
LA	Latero-Anterior
AP	Apical
SEX	Gender
HTA	Hyper-tensionY/N
DIAB	DiabetesY/N
HYPLD	HyperloipidemiaY/N
FHCAD	Familyhist.ofdisease
SMOKE	PatientsmokesY/N
CLAUDI	ClaudicationY/N
PANGIO	PreviousanginaY/N
PSTROKE	PriorstrokeY/N
PCARSUR	PriorcarotsurgY/N
CHOL	Cholesterollevel

Confidence = 1: IF 0 <= AGE < 40:0 - 1:0 <= AL < 0:2 PCARSUR = n

THEN 0 <= LAD < 50, s=0.01 c=1.00 l=2.1

IF 0 <= AGE < 40:0 - 1:0 <= AS < 0:2 PCARSUR = n

THEN 0 <= LAD < 50, s=0.01 c=1.00 l=2.1

IF 40:0 <= AGE < 60:0 SEX = F 0 <= CHOL < 200

THEN 0 <= LCX < 50 5-0.02 -1.00 -1.6

IF SEX-F HTA-n0<- CHOL < 200

THEN ORCA < 50, s-0.02 c-1.00 -1.8.

Two items in the consequent:

IF 0 4GE < 40 - 1 * 0 < - A * L < 0.2

THEN 0<- LM < 300<- LAD < 50, s-0.02 -0.89 1-1.9

IF SEX-F0CHOL < 200

THEN 0 <= LAD < 500 = RCA < 50 s = 0 -0.73 1-2.1

IF SEX-F0<- CHOL 200

THEN 0 LCX < 500<- A < 50 s-0.02 -0.73 1-1.8

Confidence >= 0:9: IF 40:0 AGE 60:01:0<-LI<0:20<-CHOL 200

THEN 0 LCX < 50, s-0.03 -0.90 1-1.5

IF 40:0 AGE 60:01:0-IL-020-CHOL 200 c

THEN 0 LCX 50, s-0 03 c- 21 -1:

IF 40 < - A * GE < 60 * 1 < - 11 < 0.2 * 0.58 - n

THEN 0 < - L * CX < 50, s - 0.01c + 0 1-1.5

IF 40:0 AGE 60:0 SEX-F DIAB-n

THEN 0 LCX 50), s-0.08 -0.92 1-15

IF HTA-n SMOKE-n 0 CHOL 200

THEN 0 LCX 50.5-0.02 -0.92 1-1.

Association Rules for Diseased Arteries:

Confidence = 1:

IF 0:2 <= SA < 1:0 HY PLPD = y PANGIO = y

THEN 70 <= LAD < 100, s=0.01 c=1.00 l=3.2

IF 60 <= AGE < 100 0:2 <= SA < 1:0 FHCAD = y

THEN not(0 <= LAD < 50, s=0.02 c=1.00 l=1.9

IF 0:2 <= IS < 1:0 CLAUDI = y PSTROKE = y

THEN not(0 <= RCA < 50), s=0.02 c=1.00 l=2.3

IF 60 <= AGE < 100:0 0:2 <= IS < 1:0 250 <= CHOL < 500

THEN 70 <= RCA < 100, s=0.02 c=1.00 l=3.2

IF 0:2 <= IS < 1:0 SEX = F 250 <= CHOL < 500

THEN 70 <= RCA < 100, s=0.01 c=1.00 l=3.2

IF 0:2 <= IS < 1:0 HTA = y 250 <= CHOL < 500

THEN 70 <= RCA < 100, s=0.011 c=1.00 l= 3.2

Two items in the consequent:

IF 0:2 <= AL < 1:1 PCARSUR = y

THEN 0 <= RCA < 50, s=0.02 c=1.00 l=1.8 THEN 70 <= LAD < 100 not(0 <= RCA < 50), s=0.01 c=0.70

l=3.9

IF 0:2 <= AS < 1:1 PCARSUR = y

THEN 70 <= LAD < 100 not(0 <= RCA < 50), s=0.01 c=0.78

l=4.4

IF 0:2 <= AP < 1:1 PCARSUR = y

THEN 70 <= LAD < 100 not(0 <= RCA < 50), s=0.01 c=0.80

l=4.5

IF 0:2 <= AP < 1:1 PCARSUR = y

THEN not(0 <= LAD < 50) not(0 <= RCA < 50), s=0.01

C=0.80 l=2.8

confidence>= 0:9:

IF 0:2 <= SA < 1:1 PANGIO = y)

THEN 70 <= LAD < 100, s=0.023 c=0.938 l= 3.0

IF 0:2 <= SA < 1:0 SEX = M PANGIO = y

THEN 70 <= LAD < 100, s=0.02 c=0.92 l=2.9

IF 60 <= AGE < 100:0 0:2 <= IL < 1:1 250 <= CHOL < 500

THEN 70 <= RCA < 100, s=0.02 c=0.92 l=2.9

IF 0:2 <= IS < 1:0 SMOKE = y 250 <= CHOL < 500

THEN 70 <= RCA < 100, s=0.02 c=0.91 l=2.9

These association rules used to identify if the patient has some symptoms of Heart Diseases or not.

CLASSIFICATION TECHNIQUES

Classification is a data mining function that assigns items in a collection to target categories or classes. The goal of classification is to accurately predict the target class for each case in the data. For example, a classification model could be used to identify loan applicants as low, medium, or high credit risks.

METHODS	ADVANTAGES	DISADVANTAGES
K-NN	1. It is easy to implement. 2. Training is done in faster manner.	1. It requires large storage space. 2. Sensitive to noise. 3. Testing is slow.
Decision Tree	1. There are no requirements of domain knowledge in the construction of decision tree. 2. It minimizes the ambiguity of complicated decisions and assigns exact values to outcomes of various actions. 3. It can easily process the data with high dimension. 4. It is easy to interpret. 5. Decision tree also handles both numerical and categorical data.	1. It is restricted to one output attribute. 2. It generates categorical output. 3. It is an unstable classifier i.e. performance of classifier is depend upon the type of dataset. 4. If the type of dataset is numeric than it generates a complex decision tree
Support Vector Machine	1. Better Accuracy as compare to other classifier. 2. Easily handle complex nonlinear data points. 3. Over fitting problem is not as much as other methods.	1. Computationally expensive. 2. The main problem is the selection of right kernel function. For every dataset different kernel function shows different results. 3. As compare to other methods training process take more time. 4. SVM was designed to solve the problem of binary class. It solves the problem of multi class by breaking it into pair of two classes such as one against-one and one-againstall.
Neural Network	1. Easily identify complex relationships between dependent and independent variables. 2. Able to handle noisy data	1. Local minima. 2. Over-fitting. 3. The processing of ANN network is difficult to interpret and require high processing .time if there are large neural networks.
Bayesian Belief Network	1. It makes computations process easier. 2. Have better speed and accuracy for huge datasets.	1. It does not give accurate results in some cases where there exists dependency among variables.

CLUSTERING

Clustering in data mining involves the segregation of subsets of data into clusters because of similarities in characteristics. This helps users better understand the structure of a data set as similar data points are put together in different groupings.

K-means clustering

Brain tumor segmentation with data mining. A group of six scientists completed their research on classifying brain tumors with the help of K-means clustering and deep learning (a subset of machine learning).

The original data sets were created from MRI scans then fed into the data mining system for preprocessing and algorithmic analysis. After passing the data down a pipeline of several statistical classifiers and geometric identification models, the system was able to differentiate between benign and malignant tumors. The resulting average accuracy turned out to be 95.62%, much higher than expected or achieved previously in similar experiments.

In order to train the model even better, scientists augmented the MRI scans with synthetic data. Deep learning algorithms require large amounts of labeled data for training, so the team took the original images and applied cropping, flipping, distortion, and noise to increase data volume.

The result was a system capable of classifying brain tumors with a phenomenal accuracy of 98.3%.

Hierarchical clustering

Big data clinical research typically involves thousands of patients and there are numerous variables available. Conventionally, these variables can be handled by multivariable regression modeling. In this article, the hierarchical cluster analysis (HCA) is introduced. This method is used to explore similarity between observations and/or clusters. The result can be visualized using heat maps and dendrograms. Sometimes, it would be interesting to add scatter plot and smooth lines into the panels of the heat map. The inherent R heatmap package does not provide this function. A series of scatter plots can be created using lattice package, and then background color of each panel is mapped to the regression coefficient by using custom-made panel functions. This is the unique feature of the lattice package. Dendrograms and color keys can be added as the legend elements of the lattice system. The lattice Extra package provides some useful functions for the work.

Density based clustering

Density-based clustering algorithms have recently gained popularity in the data mining field due to their ability to discover arbitrary shaped clusters while preserving spatial proximity of data points. In this work we adapt a density-based clustering algorithm, DBSCAN, to a new problem domain: identification of homogenous color regions in biomedical images. Examples of specific problems of this nature include landscape segmentation of satellite imagery, object detection and, in our case, identification of significant color regions in images of skin lesions (tumors). Automated outer and inner boundary segmentation is a key step in segmentation of structures such as skin lesions, tumors of breast, bone, and brain. This step is important because the accuracy of the subsequent steps (extraction of various features, post-processing) crucially depends on the accuracy of this very first step. In this paper, we present an unsupervised approach to segmentation of pigmented skin lesion images based on DBSCAN clustering algorithm. The color regions identified by the algorithm are compared to those identified by the human subjects and the Kappa coefficient, a statistical indicator of computer-human agreement, is found to be significant.

Decision trees

Decision trees in hospitals are a decision support tool that uses a tree-like structure to analyze decisions, possibilities, consequences, and measures. They may include outcomes, costs, risks, etc. Decision trees present algorithms and automate trading to offer profitable solutions.

4 Features of decision trees in healthcare call centers

From making medical decisions to empowerment, decision trees in hospitals and healthcare call centers can help in numerous ways. They are as follows:

1. Deflected calls

Decision trees in healthcare call centers deflect calls and minimize operational costs. They create single, comprehensive storage of information accessible easily.

When backed up by an efficient knowledge base it enables the seamless updating of information and provides accurate data for the searched information. This way, customers can solve their problems, thereby reducing support costs.

2. Quick call-handling

When certain customers have a problem that's beyond the self-service method, a support agent has to be involved as you can't leave the client hanging. Upon being integrated with your CRM, decision trees for healthcare call centers reduce the average handle time (AHT) and boost the first call resolution (FCL). With decision trees, you can provide solutions within a few limited steps.

3. Interactive

Healthcare call center decision trees have fun, interactive ways to engross people in the information. From step-by-step guides and visual device guides to articles and notes, any kind of content is tailor-made for the audience.

One can easily input the facts, store them safely, and enable access to information securely. Decision trees in hospitals additionally structurize information in one location.

4. Structured

Decision trees in hospitals maintain all the files that are relevant and important. They store them on a single platform where prying eyes won't wander.

Decision trees for healthcare call centers play a huge role in helping patients, their families, as well as in reducing costs and maintaining high standards. They aid customers and internal staff likewise and promote self-service and empowerment.

Decision trees in the hospital create a better work environment, lead to satisfactory solutions, and encourage growth in the right direction.

CONCLUSION

Data Mining is new technology and is still in its infancy. Applications are minimal and a very small slice of the pie has been discovered yet. Current applications are restricted to more experimental areas. Data mining should get easier and more common place every day. Health care relevant data are enormous in nature and they arrive from various birthplaces all of them not wholly relevant in structure or quality. These days, the performance of knowledge, observation of various specialists and medicinal screening data of patients grouped in a database during the analysis process, has been widely accepted. In this paper we have presented an efficient approach for fragmenting and extracting substantial forms from the heart attack data warehouses for the efficient prediction of heart attack.

Reference

Boosted Apriori: an Effective Data Mining Association Rules for Heart Disease Prediction System R.Thanigaivel and K. Ramesh Kumar Middle-East Journal of Scientific Research 24 (1): 192-200, 2016 ISSN 1990-9233 © IDOSI Publications, 2016DOI: 10.5829/idosi.mejsr.2016.24.01.22944.

International Journal of Information Sciences and Techniques (IJIST) Vol.6, No.1/2, March 2016 DOI : 10.5121/ijist.2016.6206 53 SURVEY OF DATA MINING TECHNIQUES USED IN HEALTH CARE DOMAIN Sheenal Patel and Hardik Patel.

International Journal of Bio-Science and Bio-TechnologyVol.5, No.5 (2013), pp. 241-266A survey on Data Mining approaches for Healthcare Divya Tomar and Sonali Agarwal.

International Journal of Modern Computer Science (IJMCS) ISSN: 2320-7868 (Online) Volume 6, Issue 1, February, 2018 RES Publication © 2012 Page | 19 An Overview of Data Mining - A Survey Paper.

International Journal of Innovative Research in Science, Engineering and Technology (An ISO 3297: 2007 Certified Organization) Vol. 5, Issue 8, August 2016 Copyright to IJIRSET DOI:10.15680/IJIRSET.2016.0508032 14538 A Survey of Health Care Prediction Using Data Mining Sujatha R 1 , Sumathy R 2 , Anitha Nithya R 3.