A Review on Various Data Mining Tools and Techniques for Assessment of Chronic Obstructive Pulmonary Disease (COPD)

K.Rohini
Ph.D Research Scholar, Department of Information Technology, School of Computing Sciences,
Vels University, Chennai, India.

Dr.G.Suseendran,
Assistant Professor, Department of Information Technology, School of Computing Sciences,
Vels University, Chennai, India.

Abstract – This paper is a study and review of Various Data Mining Tools and Techniques for assessment of Chronic Obstructive Pulmonary Disease. Chronic Obstructive Pulmonary Disease (COPD) is the fourth leading cause of death worldwide and represents one of the major causes of chronic morbidity. Cigarette smoking is the most important risk factor for COPD. In these patients, the airflow limitation is caused by a mixture of small airways disease and parenchyma destruction, the relative contribution of which varies from person to person. Presently, a very large amount of data stored in databases is increasing at a tremendous speed. This requires a need for new techniques and tools to aid humans in automatically and intelligently analyzing large data sets to acquire useful information. This growing need gives a view for a new research field called Knowledge Discovery in Databases (KDD) or Data Mining, which attract a attention from researchers in many different fields including database design, statistics, pattern recognition, machine learning, and data visualization. Data mining is the process of discovering insightful, interesting, and novel patterns, as well as descriptive, understandable and predictive models from large-scale data. In this paper we overviewed different tasks includes in Data mining. Data mining involves the tasks like anomaly detection, classification, regression, association rule learning, summarization and clustering.

Index Terms – Data Mining, classification, clustering, association rules; Detection; Respiratory disease; Chronic Obstructive Pulmonary Disease.

1. INTRODUCTION

The Chronic Obstructive Pulmonary Disease (COPD) is the fourth leading cause of death worldwide and represents one of the major causes of chronic morbidity. Cigarette smoking is the most important risk factor for COPD [1]. In the past this disease has been relatively neglected and unfortunately there are no current therapies that reduce the progression of the disease. Due to the enormous burden of the disease and rising healthcare cost, there is now renewed interest in the underlying cellular and molecular mechanisms and a search for new therapies, resulting in a general re-evaluation of the disease. However, despite its highly relevant health impact, there has been relatively little research into COPD and, at present, it is the most under-funded disease in relation to its global burden [2].

The Global Initiative on Obstructive Lung Disease (GOLD) has recently adopted a new definition of COPD: "a disease state characterized by airflow limitation that is not fully reversible. The airflow limitation is usually progressive and associated with an abnormal inflammatory response of the lung to noxious particles and gases" [3]. In patients with COPD, the normal inflammatory response to cigarette smoking is amplified and the typical airflow limitation is caused by a mixture of small airways disease and parenchyma destruction, the relative contribution of which varies from person to person.

The environmental factors play a vital role in many of the lung diseases. The lung capacity depends on some factors such as age, height, weight, gender, location and smoking habit . The various pulmonary function tests are present to diagnose the functioning of the lung and to detect the diseases associated with it. The Spirometry test is one of its kind. The Spirometry test data shows well you breathe in and out. Breathing in and out can be affected by lung diseases such as chronic obstructive pulmonary disease (COPD), asthma, pulmonary fibrosis and cystic fibrosis [5]. From the results of spirometry test, the patients can be classified under following three categories :

Normal

Obstructive

Restrictive

This gives two important abnormalities such as Obstructive and Restrictive lung diseases. Data mining is the exploration and analysis of large data sets, in order to discover meaningful pattern and rules . The key idea is to find effective way to

combine the computer's power to process the data with the human eye's ability to detect patterns. The objective of data mining is to design and work efficiently with large data sets. Data mining is the component of wider process called knowledge discovery from database. [4]. Data Mining is the process of analysing data from different perspectives and summarizing the results as useful information. It has been defined as "the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data".

2. DATA MINING

The definition of data mining is closely related to another commonly used term knowledge discovery [2]. Data mining is an interdisciplinary, integrated database, artificial intelligence, machine learning, statistics, etc. Many areas of theory and technology in current era are databases, artificial intelligence, data mining and statistics is a study of three strong large technology pillars. Data mining is a multi-step process, requires accessing and preparing data for a mining the data, data mining algorithm, analysing results and taking appropriate action. The data, which is accessed can be stored in one or more operational databases. In data mining the data can be mined by passing various process.

In data mining the data is mined using two learning approaches i.e. supervised learning or unsupervised learning [5].

A. Supervised Learning

In supervised learning (often also called directed data mining) the variables under investigation can be split into two groups: explanatory variables and one (or more) dependent variables. The goal of the analysis is to specify a relationship between the dependent variable and explanatory variables the as it is done in regression analysis. To proceed with directed data mining techniques the values of the dependent variable must be known for a sufficiently large part of the data set.

B. Unsupervised Learning:

In unsupervised learning, all the variables are treated in same way, there is no distinction between dependent and explanatory variables. However, in contrast to the name undirected data mining, still there is some target to achieve. This target might be as data reduction as general or more specific like clustering. The dividing line between unsupervised learning and supervised learning is the same that distinguishes discriminate analysis from cluster analysis. Supervised learning requires, target variable should be well defined and that a sufficient number of its values are given. Unsupervised learning typically either the target variable has only been recorded for too small a number of cases or the target variable is unknown.

3. ISSUES IN DATA MINING

Data mining has evolved into an important and active area of research because of the theoretical challenges and practical applications associated with the problem of discovering interesting and previously unknown knowledge from real-world databases. The main challenges to the data mining and the corresponding considerations in designing the algorithms are as follows:

1. Massive datasets and high dimensionality.
2. Over fitting and assessing the statistical significance.
3. Understandability of patterns.
4. Non-standard incomplete data and data integration.
5. Mixed changing and redundant data.

4. FUNCTIONALITIES OF DATA MINING

Data mining as a term used for the specific classes of six activities or tasks as follows:

1. Classification
2. Estimation
3. Prediction
4. Affinity grouping or association rules
5. Clustering
6. Description and visualization

The first three tasks - classification, estimation and prediction rules are examples of directed data mining or supervised learning. In directed data mining, the goal is to use the available data to build a model that describes one or more particular attribute(s) of interest (target attributes or class attributes) in terms of the rest of the available attributes. The next three tasks – association rules, clustering and description are examples of undirected data mining i.e. no attribute is singled out as the target, the main goal is to establish some relationship among all attributes [6].

A. Classification

Classification consists of examining the features of a newly presented object and assigning to it a predefined class. The classification task is characterized by the well-defined classes, and a training set consisting of reclassified examples. The task is to build a model that can be applied to unclassified data in order to classify it. Examples of classification tasks include: • Classification of credit applicants as low, medium or high risk • Classification of mushrooms as edible or poisonous • Determination of which home telephone lines are used for internet access

B. Estimation

Estimation deals with continuously valued outcomes. Given some input data, we use estimation to come up with a value for some unknown continuous variables such as income, height or credit card balance. Some examples of estimation tasks include:

• Estimating the number of children in a family from the input data of mothers' education

• Estimating total household income of a family from the data of vehicles in the family

• Estimating the value of a piece of a real estate from the data on proximity of that land from a major business centre of the city.

C. Prediction

Any prediction can be thought of as classification or estimation. The difference is one of emphasis. When data mining is used to classify a phone line as primarily used for internet access or a credit card transaction as fraudulent, we do not expect to be able to go back later to see if the classification was correct. Our classification may be correct or incorrect, but the uncertainty is due to incomplete knowledge only: out in the real world, the relevant actions have already taken place. The phone is or is not used primarily to dial the local ISP. The credit card transaction is or is not fraudulent. With enough efforts, it is possible to check. Predictive tasks feel different because the records are classified according to some predicted future behaviour or estimated future value. With prediction, the only way to check the accuracy of the classification is to wait and see. Examples of prediction tasks include: • Predicting the size of the balance that will be transferred if a credit card prospect accepts a balance transfer offer • Predicting which customers will leave within next six months • Predicting which telephone subscribers will order a value–added service such as three-way calling or voice mail. Any of the techniques used for classification and estimation can be adopted for use in prediction by using training examples where the value of the variable to be predicted is already known, along with historical data for those examples. The historical data is used to build a model that explains the current observed behaviour. When this model is applied to current inputs, the result is a prediction of future behaviour [7].

D. Association Rules

An association rule is a rule which implies certain association relationships among a set of objects (such as "occur together" or "one implies the other") in a database. Given a set of transactions, where each transaction is a set of literals (called items), an association rule is an expression of the form X Y , where X and Y are sets of items. The intuitive meaning of such a rule is that transactions of the database which contain X tend to contain Y. An example of an association rule is: "30% of farmers that grow wheat also grow pulses; 2% of all farmers grow both of these items". Here 30% is called the confidence of the rule, and 2% the support of the rule. The problem is to find all association rules that satisfy user-specified minimum support and minimum confidence constraints.

E. Clustering

Cluster analysis can be used as a standalone data mining tool to gain insight into the data distribution, or as a pre-processing step for other data mining algorithms operating on the detected clusters. Many clustering algorithms have been developed and are categorized from several aspects such as partitioning methods, hierarchical methods, density-based methods, and grid-based methods .Further data set can be numeric or categorical. Clustering is the task of segmenting a diverse group into a number of similar subgroups or clusters. What distinguishes clustering from classification is that clustering does not rely on predefined classes. In clustering, there are no predefined classes. The records are grouped together on the basis of self similarity. Clustering is often done as a prelude to some other form of data mining or modelling. For example, clustering might be the first step in a market segmentation effort, instead of trying to come up with a one-size-fits-all rule for determining what kind of promotion works best for each cluster[6].

1) General Types Of Clusters:

 Well-separated clusters: A cluster is a set of points so that any point in a cluster is nearest (or more similar) to every other point in the cluster as compared to any other point that is not in the cluster.

 Center-based clusters A cluster is a set of objects such that an object in a cluster is nearest (more similar) to the "center" of a cluster, than to the center of any other cluster. The center of a cluster is often a centroid.

 Contiguous clusters A cluster is a set of points so that a point in a cluster is nearest (or more similar) to one or more other points in the cluster as compared to any point that is not in the cluster.

 Density-based clusters A cluster is a dense region of points, which is separated by according to the low-density regions, from other regions that is of high density.

 Shared Property or Conceptual Clusters Finds clusters that share some common property or represent a particular concept.

F. Description and Visualization

Data visualization is a powerful form of descriptive data mining. It is not always easy to come up with meaningful visualizations, but the right picture really can be worth a thousand association rules since the human beings are extremely practiced at extracting meaning from visual scenes. Knowledge discovery goals are defined by the intended use of

the system. There are two types of goals: (1) verification and (2) discovery. With verification, the system is limited to verifying the user's hypothesis. With discovery, the system autonomously finds new patterns. The discovery goal is further divided into prediction, where the system finds patterns for predicting the future behaviour of some entities and description, where the system finds patterns for presentation to a user in human understandable form.

5. DATAMINING TOOLS

We have more than 40 data mining tools

Widely used data mining tools under medical domain are as follows:

ORANGE : Orange is an open source data visualization and analysis tool, where data mining is done through visual programming or Python scripting. The tool has components for machine learning, add-ons for bioinformatics and text mining and it is packed with features for data analytics.

WEKA : Weka is a collection of machine learning algorithms for data mining tasks. The algorithms can either be applied directly to a dataset or called from your own Java code. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization.

RAPIDMINER : RapidMiner is a software platform developed by the company of the same name that provides an integrated environment for machine learning, data mining, text mining, predictive analytics and business analytics. It is used for business and commercial applications as well as for research, education, training, rapid prototyping, and application development and supports all steps of the data mining process including data preparation, results visualization, validation and optimization

R: R is a programming language and software environment for statistical computing and graphics supported by the R Foundation for Statistical Computing. The R language is widely used among statisticians and data miners for developing statistical software and data analysis. Polls, surveys of data miners, and studies of scholarly literature databases show that R's popularity has increased substantially in recent years.

6. CONCLUSIONS

Data mining involves extracting useful rules or interesting patterns from huge historical data. Many data mining tasks are available and each of them further has many techniques. Data mining is an interdisciplinary, artificial intelligence, integrated database, machine learning, statistics, etc. Data mining is a large number of incomplete, noisy, fuzzy, random application of the data found in hidden, regularity which are not known by people in advance, but is potentially useful and ultimately understandable information and knowledge of non-trivial process This paper discusses about some issues in Data Mining and activities and Tools used to perform Data mining task.

REFERENCES

[1] Amandeep Kaur Mann, Navneet Kaur, *Survey Paper on Clustering Techniques, IJSETR,2278 – 779.*
[2] Pavel Berkhin, A *Survey of Clustering Data Mining Techniques*, pp.25-71, 2002.
[3] Oded Maimon, Lior Rokach, *Data Mining AND Knowlwdge Discovery Handbook*, Springer Science + Business Media.Inc, pp.321-352, 2005.
[4] Han, J., Kamber, M., *Data Mining Concepts and Techniques*, Morgan Kaufmann Publisher, 2001
[5] K.Kameshwaran, K.Malarvizhi, *Survey on Clustering Techniques in Data Mining*, IJCSIT, Vol. 5, 2014, 2272-2276
[6] Aastha Joshi, Rajneet Kaur, *A Review: Comparative Study of Various Clustering Techniques in Data Mining*, IJARCSSE, Vol. 3, 2013, 2277 128X
[7] Manish Verma, Mauly Srivastava, Neha Chack, Atul Kumar Diswar, Nidhi Gupta, *A Comparative Study of Various Clustering Algorithms in Data Mining,, International Journal of Engineering Reserch and Applications (IJERA)*, Vol. 2, Issue 3, pp.1379-1384, 2012.
[8] Pradeep Rai, Shubha Singh, *A Survey of Clustering Techniques*, International Journal of Computer Applications, 2010.