A Survey on Process of Data Mining With Techniques

B.SOWMIYA
SIVET COLLEGE, UNIVERSITY OF MADRAS

Abstract- *Data mining plays a major role in research area and used by the biologists to statisticians and computer scientists as well. Data mining is helpful in acquiring knowledge from large databases, data warehouses and data marts. The main purpose of this paper is to discuss the process of data mining steps with techniques.*

Keywords- Data mining, knowledge discovery in databases, techniques in data mining, clustering, classification.

I. INTRODUCTION

data mining is an iterative and semi automatic process of discovering knowledge from the existing large data sets . Data mining is a combination of computer
science and statistics with the goal to extract information from a large data set and transform the information into structure for further use.

Data mining is the analysis step of the KDD or "knowledge discovery in databases" process. Also it involves database and data management aspects, data preprocessing model . Data mining focuses on using specific machine learning and statistical models to predict the future and discover the patterns among data.

The techniques covered include association rules, sequence mining, decision tree classification, and clustering.

The rest of the paper is organized as follows. In section. In Section II, discussing the characteristics of mined data and on section III, give the short explanation about the steps in data mining. In section IV, the techniques of data mining are given. Finally conclusion is presented in section V of the paper and references that are made are presented in Section VI

II. CHARACTERISTICS OF THE MINED DATA

This section discusses the crucial characteristics of data which is mined.

(a) **Valid**:

It is required that the patterns, rules, and models that are discovered are valid not only in the data samples already examined, but are generalizable and remain valid in future new data samples. Only then can the rules and models obtained be considered meaningful.

(b) **Novel**:

It is desirable that the patterns, rules, and models that are discovered are not already known to experts. Otherwise, they would yield very little new understanding of the data samples and the problem at hand.

(c) **Useful**:

It is desirable that the patterns, rules, and models that are discovered allow us to take some useful action. For example, they allow us to make reliable predictions on future events.

(d) **Understandable**:

It is desirable that the patterns, rules, and model that are discovered lead to a new insight on data samples and the problem being analysed.

III. STEPS IN DATA MINING

The general data mining process is shown in Figure 1. It comprises the following steps , some of which are optional depending on the problem being analysed.

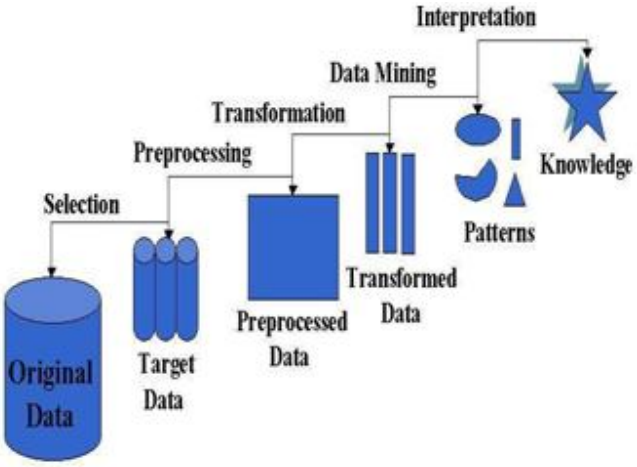

Fig 1: The data mining process

(1) **Understand the application domain**:

data mining requires a proper understanding of the application domain to get the outcomes desired by the user. prior knowledge is added to maximize the chance of success.

(2) **Collect and create the target dataset**:

Data mining relies on the data set being analyzed. Therefore, the collection of a dataset that captures all the possible situations that are relevant to the problem being analyzed is crucial.

(3) **Clean and transform the target dataset**:

Raw data contain many errors and inconsistencies, such as noise, outliers, and missing values. An important element of this process is the de-duplication of data records to produce a non-redundant dataset. For example, in collecting informa- tion from public sequence the same sequence may be recorded multiple times in the public sequence databases, this may lead to unnormalized and missing values and these need to be dealt with properly.

(4) **Select features, reduce dimensions**:

Even after the data have been cleaned up in terms of eliminating duplicates, inconsistencies, miss- ing values, and so on, there may still be noise that is irrelevant to the problem being analyzed. These noise attributes may confuse subsequent data mining steps, produce irrelevant rules and associations, and increase computational cost. It is therefore wise to perform a dimension reduction or feature selection step to separate those attributes that are pertinent from those that are irrelevant. This step is typically achieved using statistical or heuristic techniques such as Fisher criterion ,Wilcoxon rank sum test , principal component analysis , entropy analysis, etc.

(5) **Apply data mining algorithms**:

Now we are ready to apply appropriate data mining algorithms,|association rules discovery, sequence mining, classification tree induction, clustering, and so on|to analyze the data.Some of these algorithms are discussed in later sections.

(6) Interpret, evaluate, and visualize patterns:

After the algorithms above have produced their output, it is still necessary to examine the output in order to interpret and evaluate the extracted patterns, rules, and models. It is only by this interpretation and evaluation process that we can derive new insights on the problem being analyzed.

IV. DATA MINING TECHNIQUES

Data mining relies on techniques like Association, classification,clustering, prediction, sequential pattern mining, categorization, estimation, and visualization, etc. Classification assists with identifying associations and clusters, and separates subjects under study. E.g., education institutions can use classification comprehensively to analyze student's characteristics.

Categorization applies rule induction algorithms to handle categorical outcomes. Estimation includes predictive functions or likelihood deals with continuous outcome variables. Estimation and classification use unsupervised orsupervised modeling techniques. Visualization uses interactive graphs to demonstrate mathematically induced data and scores, and is much more sophisticated than traditional bar charts or pie charts.

An algorithm is a specific, mathematically driven data mining function, such as a neural network, classification and regression tree (C&RT), or K-means. Data mining techniques including algorithms such as clustering, classification, regression, neural networks, association rules, decision trees, some of them have been applied successfully in the educational area.

E.g., methods for hierarchical data mining and longitudinal data modeling have been applied into EDM.

Clustering

The goal of clustering is to group similar set of objects together, splitting the full data set into clusters sets. By using clustering techniques we are able to further identify dense and sparse regions in object space, and discover overall distribution pattern and correlations among data attributes. Cluster analysis is not a specific algorithm, but the general task to be solved.Thus clustering itself may be formulated as a multi objective optimizatio problem. Clustering algorithms may be used for organizing data, categorize data for model construction and data compression, outlier detection, etc.

Classification

The goal of classification is to organize and categorize data in distinct classes.classification is used to predict values for some variables. This algorithm frequently

employs the decision tree or neural network-based classification algorithms. In classification test, data are used to estimate the accuracy of the classification rules. If Data-Mining Research in Education the accuracy is acceptable then rules are able to be applied into new data. Some popular classification methods include decision trees, logistic regression (for binary predictions) and support vector machines. There are several classification models. Some of the common classificat ion models are decision trees,neural networks, genetic algorithms, support vector machines, Bayesian classifiers. The application includes credit risk analysis, fraud detection, banking and medical application, etc.

Association rule mining

Association rule mining is a rule- based ,well-researched method for discovering interesting relations between variables in large databases. Association rules are if-then statements that shows the probability of relationships between data items in large data sets of various databases. A classic example association rule mining refers to relationship between diapers and beers. About 5500 transactions(2.75) include purchase of beer.of those , about 3,500 transactions, 1.75% include purchase of diapers and beer.

Regression

Regression analysis is a statistical processes for estimating relationship between independent variables and dependent variables. Independent variables are known attributes and response variables are able to predict what we want. Regression is a technique used to predict a range of numeric values.For example regression is used to predict the cost of a product or service, given other variables.

Neural Networks

Neural networks is an adaptive system that changes its structure during learning phase.. The Neural networks have the ability to extract meaningful and useful patterns and trends from the complex data. It is applicable to real world problems especially in case of industry. . Artificial neural network (ANN) learns by example. ANN is configured for specific application as classification, pattern recognition etc. through a learning process. It may also be used for three dimensional object recognition, hand-written word recognition, face recognition, etc. Neural networks have the drawback of not explaining the derived results. Another problem is that it suffers from long learning times. As the data grows, the situation becomes worse for that problem.

Decision Trees

Decision tree is tree-shaped structures(includes root node, branches and leaf nodes) which represents decisions sets.Decision trees start with a root which then branches off into a number of solutions. Specific decision tree methods include classification and regression trees (CART) and Chi Square Automatic Interaction Detection.

Nearest Neighbour Method

Nearest neighbour method, also called the k-nearest neighbour technique uses classification and regression. each record in the dataset based on a combination of different classes of the k record(s), which is similar to that in a historical dataset (where k >=1). To choose the appropriate algorithms, researchers need design the data and align it with the desired output.

Support Vector Machines

Support vector machines (SVM) are supervised learning models with associated learning algorithms that analyse data used for classification and regression analysis.
SVMs select the plane which maximizes the margin separating the two classes. The margin is defined as the distance
between the separating hyperplane to the nearest point of A, plus the distance from the hyperplane to the nearest point in B, where A and B aretwo linearly separable sets. SVM has been used in many applications including face detection, handwritten character and digits recognition,speech recognition, image and information retrieval.

Genetic Algorithms

Genetic algorithms is based on Darwin's theory of evolution. A population of the individual with possible solution to a problem is created initially at random. Then the crossover is done by combining pairs of individuals to produce offspring of next generation. A mutation process is used to modify the genetic structure of some members of new generation randomly. The algorithm searches for a solution in the successive generation. When an optimum solution is foundor some fixed time is elapsed, the process comes to an end. Genetic algorithms are widely used in problems where optimization is required.Genetic algorithms repeatedly modifies population of individual solutions.

V. CONCLUSION

In this paper we have discussed the data mining with process steps, characteristics of mined data and various data mining techniques Such as classification, clustering,

regression, Decision trees, Genetic algorithms, neural networks,etc…

REFERENCES

[1] Sadiq Hussain , "Survey on Current Trends and Techniques of Data mining Research", London Journal of Research in Computer Science and Technology, 2017, Volume 17, Issue 1 Compilation 1.0
[2] Adam Baba, Gouse Pasha, Shaik Althaf Ahammed, S. Nasira Tabassum, "Introduction to Neural Networks Design Architecture", International Journal of Scientific & Engineering Research Volume 4,Issue 2, Februry 2013, ISSN 2229-5518.
[3] Arun K Pujari, Data Mining Techniques,University Press, 2013.
[4] H. Kargupta and A. Joshi, "Data Mining to Go: Ubiquitous KDD for Mobile and Distributed Environments", KDD-2001, San Francisco, August 2001.
[5] Jiawei Han and Micheline Kamber, Data Mining: Concepts and Techniques, Morgan Kaufmann Publishers, 2003.