Data Mining Techniques for the Knowledge Discovery

Tarika Verma[1], Dr. Chhavi Rana[2]

[1.] Department of Computer Science and Engineering, UIET,
M.D. University, Rohtak- 124001, Haryana, India
Email: tarika.verma@yahoo.co.in

[2.] Assistant Professor, Department of Computer Science and Engineering, UIET,
M.D. University, Rohtak - 124001, Haryana, India
Email: chhavi1jan@yahoo.com

Abstract: **The knowledge discovery process is done using the data mining techniques by transforming the raw data from various sources into meaningful patterns and further interpreting those into useful information. Various data mining algorithms like K-Means or a Priori may be used for the purpose of extraction of meaningful patterns or trends in the given data. This review paper contains various data mining techniques in a comprehendible way.**

Keywords: Association, Classification, Clustering, Data Mining Techniques, Decision Trees, Knowledge Discovery, Neural Networks

I. INTRODUCTION

Basically, the main goal of knowledge discovery and data mining is to discover the unrevealed information from the provided data. This information may further be used for many purposes, depending upon the nature of the extraction for the same data sets.

The hidden facts unveiled by the data mining techniques for instance may be used by a business organization for the purpose of increasing the sale of a particular product or for gaining the information regarding the behaviour of its customers towards their existing product. Similarly the medical organizations may use the data mining for better diagnostics of the known and unknown diseases etc. Also the data mining may be used to find the unknown data records for a certain data collection or to correct the erroneous data sets. Defining data mining in simple terms, it is used to detect the useful trends or patterns in the given database.

Data Mining Techniques are the way to achieve the goal of the data mining or we can say to unveil the useful hidden facts and patterns in a database. Using these techniques, the database is scanned for the similarities and dissimilarities in order to find the relationship between the existing and non existing data objects. [1]Various types of data mining techniques are available:

- Classical Mining Techniques: Association, Classification and Clustering
- Modern Techniques: Decision Trees, Neural Networks

II. Classical Mining Techniques

A. Association

In association the relation between the various data objects is found and a particular pattern is found out. Hence this technique is also known as Relation Data Mining Technique.

This technique is similar to the analysis which is done by the customers in market basket selection. [2] This helps in identifying the frequently bought together products specifically for the prediction of those products which will be present if certain other product is present in the basket. Further this information may be used to know more about a certain customer. This may help in increasing the sale of the associated products when referred to a specific customer. This may thus be said: It is more likely that certain products occur together in the selected items in a basket. Thus association rule helps us in predicting the co- occurrence of the certain products. [3]

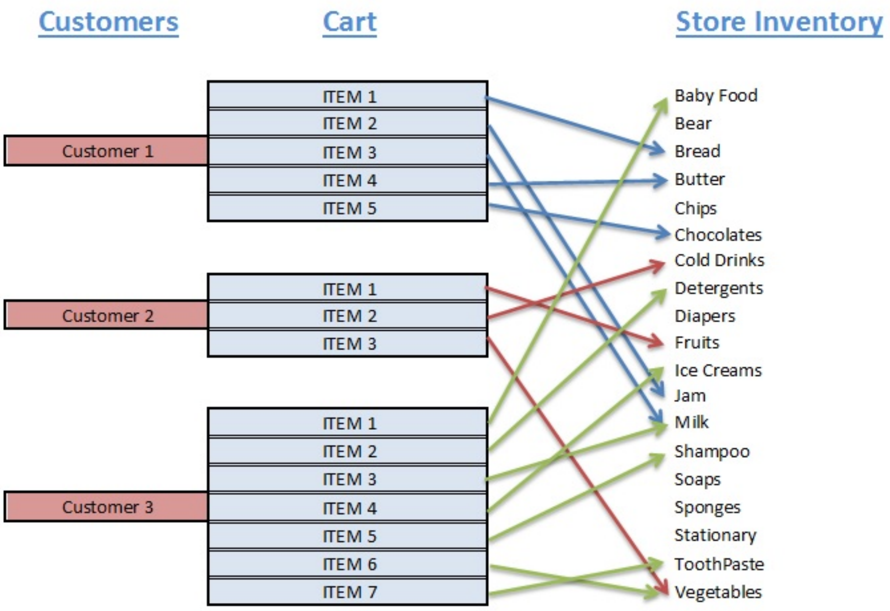

Fig 1. Customers associated to products bought together

The association rules can be applied only when the data of interest is collectively taken under consideration. It can't be used when the data is in scattered form. As the result of applying the association rules on the given collection of data, we get the corresponding pattern in the present data objects or the events.

B. Classification

Classification classifies each data item into one of a predefined set of classes or groups. [4]

For the classification [5]:

1) Model Construction: The classes or groups are predicted and class labels are specified. Using the training data and the classification algorithms, the classes are created
2) Using Model for Prediction: The unseen data is classified based on the class model.

The classification model is denoted using classification (or grouping) rules and using decision trees.

This is a type of supervised learning because the training data sets are firstly given labels which indicate the class names and then the unseen data is classified as per the training set. [6]

C. Clustering

In clustering the data items with similar properties are divided into clusters or groups. Here we have unsupervised learning as the classes are not pre defined as in case of classification technique.

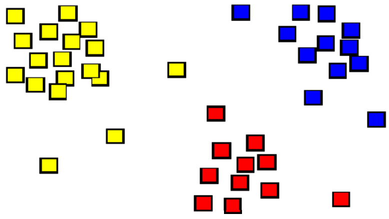

Fig 2. Clustering [7]

Clustering causes generalization of items in the sample space and thus the fine details may be lost. But with this generalization the simple groups of data objects or items is obtained, thus in spite of the abstraction of finer properties, we get, some simple clusters having some feature(s) in common. [8]

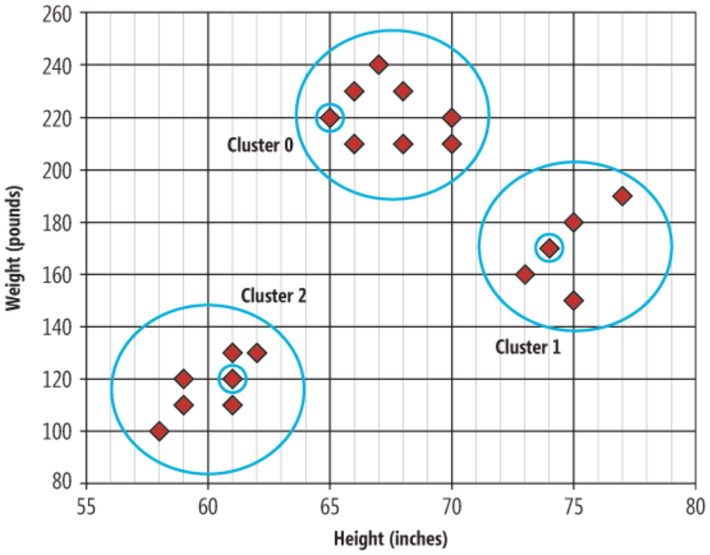

Fig 3. K-means algorithm output having clustered data and respective centroids for sample data [17], k=3

In its model it forms the clusters of its data items which correspond to the unseen patterns or features.

Classification of Clustering Algorithms [9]:

1) Hierarchical Clustering Methods: Agglomerative Algorithms (The bottom-up approach for making clusters) and Divisive Algorithms (The top-down approach for making clusters)
2) Partitioning Relocation Methods: Probabilistic Clustering Method, K-medoids Clustering Method and K-means Clustering Methods

In the Hierarchical clustering methods a hierarchical tree of the clusters or groups is formed which is also called a dendrogram.

In the data partitioning algorithms the given data to be clustered is firstly divided into several subsets. Then some greedy heuristics is applied iteratively. This gradually increases the quality of clusters finally. [10]

Practically clustering plays a formidable role in most of the applications of data mining like in text mining, medical diagnostics, business purposes and many others.

III. MODERN MINING TECHNIQUES

A. Decision trees

Decision Tree is a “divide-and-conquer” approach, from a given set of self-reliant instances, which give rise to a learning problem. In this technique, the root node of the decision tree represents a simple problem statement or a condition that has one or more solutions. Each solution then gives rise to a set of problems or conditions that guides us to determine the final decision based on it. [11]

Here all the nodes belonging to a decision tree involve testing of one or more properties, or use some function of these properties.[12] Leaf nodes of the decision tree give a classification that applies to all the instances that come in the path to reach that particular leaf node. Thus when an unseen instance is tested, it is routed from the root to the node on the next tree level according to its properties tested in consecutive nodes, and on reaching a leaf that instance is classified as per the classification of the leaf. [13]

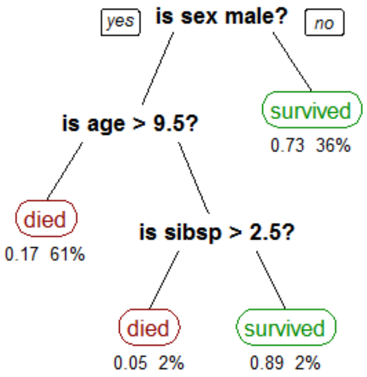

Fig 4. Decision tree to show the survival of Titanic passengers ("sibsp" = no. of spouses or siblings aboard) [14]

The main advantage of a decision tree is that it is easy to comprehend for users and thus it is used in a vast area of data mining applications. Also they state the problem in the form of IF-Then rules and thus making the representation meaningful.

B. Neural Networks

Neural Networks (NN) is a recognized technology which resembles the Biological Neural network. It consists of a basic unit called neuron. There are numerous interconnected neurons that process the information at an extremely fast speed. One neuron is connected with other neuron using a numerical value called weight. The total input signal received by a neuron is in the form of combined inputs to that neuron from the other neurons.

Each neuron receives signals from connected neurons and the combined input signal is calculated. The total signal inputted to a neuron j is

$$u_j = \sum w_{ij} * x_i,$$

where x_i is the input signal from any neuron i and w_{ji} is the connection weight between the i and j neurons. [15]

The neuron is fired only when the combined input is greater than a threshold value. [16] The arrangement of neurons is in the form of layers. Basically it contains one input layer, zero or more hidden layer and one output layer in its layered network. After defining the network architecture, the network must be trained. For instance, in back-propagation networks, a pattern is given to the input layer of the network and a final output of the network is calculated at the output layer. The output is compared with the required outcome and the errors are propagated backwards by adjusting the connection weights. This process is iterated until we reach an acceptable error rate. [17]

The neural networks may be used for a wide variety of applications. [18] They are used in detecting the credit card fraudulent use, clustering, prototype creation, feature extraction and many others.

REFERENCES

[1] Kurt Thearling, “An Overview of Data Mining Techniques”, Thearling.com

[2] Hegland, Markus. "Data mining techniques." Acta Numerica 2001 10 (2001): 313-355.

[3] Martin Brown, “Data Mining Techniques”, IBM.com, 2012

[4] Han, Jiawei, Jian Pei, and Micheline Kamber. Data mining: concepts and techniques. Elsevier, 2011.

[5] By Cluster-2.gif: hellisp derivative work: Wgabrie (Cluster-2.gif) [Public domain], via Wikimedia Commons

[6] Berkhin, Pavel. "A survey of clustering data mining techniques." Grouping multidimensional data. Springer Berlin Heidelberg, 2006. 25-71.

[7] Witten, Ian H., et al. Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann, 2016.

[8] By Stephen Milborrow (Own work) [CC BY-SA 3.0 (http://creativecommons.org/licenses/by-sa/3.0) or GFDL (http://www.gnu.org/copyleft/fdl.html)], via Wikimedia Commons

[9] Kirkos, Efstathios, Charalambos Spathis, and Yannis Manolopoulos. "Data mining techniques for the detection of fraudulent financial statements." Expert systems with applications 32.4 (2007): 995-1003.

[10] Han, J., & Camber, M. “Data mining concepts and techniques” San Diego, USA: Morgan Kaufman., 2000 : 62-65

[11] James McCaffrey, “Data Clustering - Detecting Abnormal Data Using k-Means Clustering”, MSDN Magazine Vol 28 No. 2, February 2013

AUTHOR PROFILE

Tarika Verma is currently pursuing Master in Technology degree at MD University, Rohtak, Haryana, India. She has also got TEQUIP World Bank scholarship for the meritorious students and has been teaching under the same scheme since last two years. She has been interested in the field of Database Management, Data Analytics, and Big Data Analysis. She has attended conferences and also has presented papers related to this field.

Dr. Chhavi Rana Balhara has an experience of over 10 years teaching Data Mining and Web Development subjects at various engineering institutes. She has been interested in the area of Data Mining research from the past 8 years attending around 25 conferences and presenting papers related to this field. Also, she has published 30 papers in reputed journals including Springer and Elsevier. Besides data mining, her research interests also include information management, information retrieval, and ICT. She has supervised 20 M.Tech thesis and currently supervising 4 M.Tech thesis and 4 Ph.D. students. She has also won DST Travel grant twice to present the paper in USA and Spain as well as TEQUIP World Bank travel grant to present a paper in University of Sydney, Australia. She has also been a reviewer on IEEE Transaction’s on Systems, Man and Cybernetics: Systems, Artificial Intelligence Review, Springer as well as Inderscience Publishers. She has also published 4 books on her research work.