Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X
IMPACT FACTOR: 6.017

IJCSMC, Vol. 6, Issue. 7, July 2017, pg.227 – 231

Data Mining Techniques Methods Algorithms and Tools

Nelofar Rehman

Assistant Professor, PG Department Of Computer Science, SSM College Of Engineering and Technology, Kashmir, India

nelofar.rehman@gmail.com

……..……..***……..……

Abstract: Data mining is the practice of examing large pre-existing database in order to generate new information. It is a powerful technology with great potential to help companies focus on most important information in their data warehouses. Data mining tools predicts future trends and behaviour allowing business to make proactive, knowledge driven decisions. Data mining tools can answer business questions that traditionally were too time consuming to resolve.

1.Introduction

Data mining is the process of extraction hidden knowledge from volumes of raw data through use of algorithm and techniques drawn from field of statistics, machine learning and data base management system. Data mining also called knowledge discovery in large data enables firm and organization decisions by assembling, accumulating, analyzing and accessing corporate data. It uses variety of tools like query and reporting tools, analytical processing tools and decision support system(DSS) tools.

2.Key Properties Of Data Mining

- Automatic discovery of patterns
- Prediction of likely outcomes
- Creation of actionable information
- Focus on large data sets and database

Data mining can answer questions that cannot be addressed through simple query and reporting techniques.

2.1 Automatic Discovery Of Patterns

Data mining is accomplished by building models. A model uses algorithm to act on right models of data. The notion of automatic discovery refers to execution of data mining models. Data mining tools can be used to mine data on which they are built but most of the models are generalized to new data. The process of applying a model to new data is known as scoring.

2.2 Prediction

Many forms of data mining are predictive example a model might predict income based on education and other demographic factors. Prediction probabilities are known as confidence.

Some forms of predictive data mining generate rules which are conditions that imply a given outcome example a rule might specify that a person who has a bachelors degree and lives in a certain neighbourhood is likely to have an income greater than regional average.

2.3 Grouping

Other forms of data mining identify natural groupings in data example a model might identify the segment of population that has income within specified range that has a good driving record and that leases a new car on yearly basis.

2.4 Actionable Information

Data mining can derive information from large volumes of data example a town planner might use a model that predicts income base on demographies to develop a plan for low income housing. A car leasing agency might use model that identifies customer segments to design a promotion targeting high value customers.

3. Data Mining Process

Data mining process has iterative nature. Data mining project doesnot stop when a particular solution is deployed. The result of data mining trigger new business questions which in turn can be used to develop more focussed models.

3.1 Problem Definition

This initial phase of data mining project focusses on understanding the project objectives and requirements example our business problem might be "How can sell more of my products to customers"? we might translate this into data mining problem such as "which customers are most likely to purchase the product?"A model that predicts who is most likely to purchase the product must be built on data that describes the customers who have purchased the product in past. Before building the model we must assemble the data that is likely to contain relationship between customers who have purchased the products and customers who have not purchased the product. Customer attributes might include age,no.of children,years of residence.owners and so on.

3.2 Data Gathering and Preparation

The data understanding phase involves data collection and exploration. As we take a closer look at data we can determine how well it addresses the business problem. We might decide to remove some of data or add additional data. This is also the time to identify data quality problems and to scan for patterns in data.

3.3 Model Building And Evaluation

In this phase we select and apply various modelling techniques and callibrate the parameters to optimal values. If the algorithm requires data transformations we will need to step back to previous phase to implement them. At this stage of the project iot is time to evaluate how well the model satisfies originally stated business goal.

3.4 Knowledge Deployment

Knowledge deployment is the use of the data mining within a target environment. In the deployment phase insight and actionable information can be derived from data. Deployment can involve scoring the extraction of model details or the integration of data mining within applications data warehouse infrastructure or query and reporting tools.

4. Techniques Of Data Mining

To analyse large amount of data, data mining came into picture and is also known as KDD process. To complete process various techniques are deployed so afra. Data mining uses already build tools to get out useful hidden patterns trends and predictions of future can be obtained using techniques.

4.1 Classification

Classification is one of the data mining technique which is useful for predicting group membership for data instance example a classification model can be used to identify loan applicants is low, medium or high credit tasks.

4.1.1 Methods

a.Decision tree induction: A decision tree is a structure that includes a root node branches and leaf node. Each internal node represents a test on attribute ,each branch denotes the outcome of test each leaf node holds a class label. The top[most node in tree is root node. The main goal is to predict the output of continuous attribute but data mining is less appropriate for estimating tasks.

b.Rule: It is represented by set of IF-THEN riles. First of all how these rules are examined and next is how these rules are build and can be generated from data. Expression for rule is

IF condition THEN conclusion

4.2 Clustering

By examing one or more attributes or classes we can group individual pieces of data together to form a structure opinion. At a simple level ,clustering is using one or more attributes as our brain for identifying a cluster of corelating results.

4.2.1 Methods

a. Partitioning method: Suppose we are given a database of n objects and partitioning method constructs k partitions of data. Each partition will represent a cluster and k<=n. It means that it will classify data into K groups which satisfy

- each group contains atleast one object
- each object must belong to exactly one group

b.Hierarchial methods: This method creates a hierarchial decomposition of given set of data objects. We can classify hierarchial methods on basis of how hierarchial decomposition is formed. There are two approaches agglomerative approach and division approach.

4.3 Regression

Regression is used to predict continuous and numerical target. It predicts no., sales, profit, square footage temperature rates. It is base on training process. It estimates value by comparing already known and predicted value.

4.3.1 Methods

a. Linear regression: It is used where relationship between target and predictor can be represented in straight line.

$$y=p1x+p2+e$$

b.Non-linear regression: It is used where relationship between target and predictor can't be represented in straight line.

5.Data Mining Algorithms

category	algorithm
• classification	c4.5,support machine vector
• clustering	Kmean algo
• Hierarchial	BIRCH,CURE

6.Data Mining Tools

- Rapid Miner:It claims to be "world leading open source system for data and text mining".Rapid analytics is a server version of that product.
- Mahout:This Apache project offers algo for clustering class and batch based calloborative filtering that run on top of Hadoop.
- Orange:This Project hopes to make data mining fruitful and fun for both novices and experts. It offers a wide variety of visualization plus a toolbox of more than 100 widgets.
- weka:short for "waikato environment for knowledge analysis".Weka offers a set of algo for data mining that we canapply directly to data or use in another java application.
- DataMelt :It can do mathematical computation data mining statiscal analysis and data visualization.

7.Conclusion

This paper presents detailed description odf data mining and its techniques and best methos and algos for techniques. Today all IT professionals engineers and researchers are working on big data. Big data is term of concerning large volumes of complex data sets .The high performance computing paradigm is required to solve the problem of big data.

References

[1].xingqan hu,lan Davidson,'knowledge Discovery and Data Mining:challenges and Realities",ISBN 978-1-59904-2521 Hershey ,New York,2007

[2].Grabmeir.J,Rudolph,'Technique of clustering Algorithms in Data Mining ",Data Mining and Knowledge Discovery,2002

[3] Ali,showkat and Kate A.Smith"On learningalgo selection for claasification"Applied soft completing 2006.

[4].https://docs.oracle.com/cd/B28359-01/datmine.111/b28129/process.htm

[5].https://en.wikipedia.org/wiki/data-mining

[6].opensourceforu.com/2017/03/top-10-opensource-data-mining-tools

BIOGRAPHIES

Nelofar Rehman is Assistant Professor at SSM College Of Engineering And Technology in PG Department Of Computer Sciences from University Of Kashmir, J&K, India. Her field of interest is Data Analysis, Artificial Intelligence and Algorithms.

