Vol 4 Issue 12 June 2015

ISSN No :2231-5063

International Multidisciplinary Research Journal

Golden Research Thoughts

Welcome to GRT

RNI MAHMUL/2011/38595 **ISSN No.2231-5063**

Golden Research Thoughts Journal is a multidisciplinary research journal, published monthly in English, Hindi & Marathi Language. All research papers submitted to the journal will be double - blind peer reviewed referred by members of the editorial board.Readers will include investigator in universities, research institutes government and industry with research interest in the general subjects.

International Advisory Board

Editorial Board

Address:-Ashok Yakkaldevi 258/34, Raviwar Peth, Solapur - 413 005 Maharashtra, India
Cell : 9595 359 435, Ph No: 02172372010 Email: ayisrj@yahoo.in Website: www.aygrt.isrj.org

International Recognized Double-Blind Peer Reviewed Multidisciplinary Research Journal
Golden Research Thoughts
ISSN 2231-5063 Impact Factor :3.4052(UIF)
Volume - 4 | Issue - 12 | June - 2015 Available online at www.aygrt.isrj.org

DATA MINING : CHOOSING ALGORITHMS

Sathish. S. N
Project Manager, Infosys Limited, Mysore, India.

Short Profile

Sathish. S. N is a Project Manager at Infosys Limited of Mysore, India.

ABSTRACT:
This paper aims to describe the different types of Algorithms in Data Mining with examples and helps to choose the algorithm based on what the business requirements are.

KEYWORDS
Data Mining, ChoosingData Mining Algorithms.

INTRODUCTION

Data Mining:

Data mining is the technique of extracting the meaningful information from large and mostly unorganized data banks. It is the process of performing automated extraction and generating predictive information from large data banks. This process of extracting the meaningful information

from the large data banks is otherwise called as knowledge discovery. Data mining is the integration of various techniques from multiple disciplines such as statistics, machine learning, pattern recognition, neural networks, and image processing and database management systems

Data Mining Algorithm

A data mining model is created by making use of data mining algorithm(s). These algorithms analyze for specific behaviors in the data set and define parameters of the mining model. These parameters are applied for the whole data set to get statistics which will help to predict, correlate and summarize.

Some Types of Data Mining Algorithm

Some types of Data mining algorithms are:-

1.Clustering Algorithm

The cases with similar behavior are grouped together by using iterative techniques. These techniques help in exploring data, identifying deviation in data and making predictions. Certain relationships which one might not derive normally are identified by clustering models. E.g. in a car company customers buying car in a particular range are grouped together and in a super market customers buying similar products are grouped together.

Steps involved in clustering algorithm working are as follows:-

1. Identifies dataset relationships.
2. Creates an order of clusters depending on identified relationships.
3. Groups cluster points on the graph which combine all cases in the dataset and demonstrate the relationships that the algorithm determines.
4. Calculates the extent to which these clusters represent the grouping of points.
5. Iterates the above steps given until the outcome cannot be further improved.

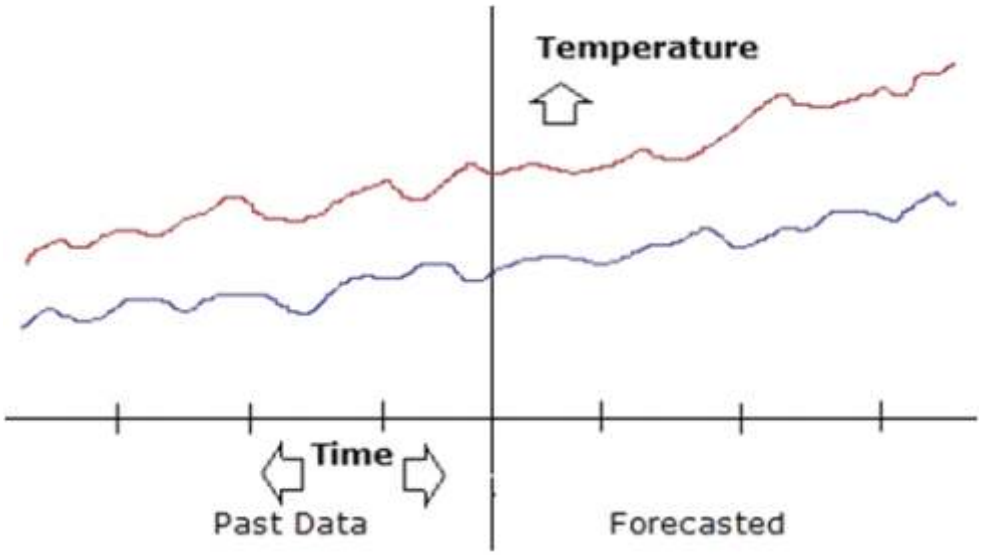
Lets us see how these steps work by using the car company example.

Step 1: Customers buying different cars from the company are identified.
Step 2: Cars are classified into different clusters based on their price.
Step 3: Cars with closer range in price are grouped together and so are the customers buying those cars.
Step 4: Calculates by how far classifying of cars and the customers who buy them into groups (clusters) are accurate based on the price of the cars.
Step 5: Above steps are iterated until the above result cannot be further improved.

2.Time Series Algorithm

Time Series algorithm provides regression algorithms for forecasting continuous values. This algorithm does not need additional columns of new data with as input to forecast a trend. New information can be added to the model when one makes a prediction and this new information can be automatically taken care of in the trend analysis.

Let us take an example of predicting global climate change based on the average temperature of the earth for each year's winter and summer season for the past century. So here the century becomes the period and temp for each year (for winter and summer) becomes the case series. The algorithm calculates the temperaturefor the future century based on the temperatures of the past century as given in the figure below (winter in blue and summer in red):



Microsoft Time Series algorithm can carry out cross prediction. Suppose there are two different series which are related. You can use this algorithm to predict the result of one series based on the characteristics of the other. One more use from a cross prediction is that a general model can be created which can be used in multiple series.

3.Naive Bayes Algorithm

This algorithm employs Bayes theorem. This algorithm's main advantage is when mining models have to be generated quickly in order to figure out relationships between input and predictable columns. Compared to rest of the algorithms, Naive Bayes algorithm is less precise and not that computation concentrated. Hence it is used for initial examination of data.

Let us consider a scenario where this algorithm will be useful. A baby product company wants to promote their products by giving free feeding bottles to its customers. However the company wants to minimize the cost by sending free bottles only to those families who have expectant mothers or babies under 3 years of age. In case Naïve Bayes algorithm can quickly give the results on which the company can send the feeding bottles to those customers who are likely to purchase their products.

For each feasible state of the predictable column Naïve Bayes algorithm estimates probability of various states of each input column. It creates mining models making use of Predictive Model Markup Language (PMML), supports drill through and backs the use of OLAP mining models. Data mining dimensions are not supported by this algorithm.

4.Association Algorithm

Recommendation engines make use of Association Algorithm. Depending on the products customers have bought, a recommendation engine recommends same or other items to customers. It can also be used for market basket analysis.

Built on datasets which contain identifiers both for individual cases and items that contain cases, association models consist of an order of item sets (group of items in a case) and rules which specify how items are band together within cases. Based on the items in the shopping basket of a customer,rules that algorithm figures out is used to forecastwhat the customer possibly would purchase in future.

Let us take the case of an airlines company. The company wants to know which all air travel routes customer is likely to travel in the future. Based on number of customers who are likely to travel by using the same route in future, the company can decide to start a new passenger aircraft on this route. Thus association algorithm can be helpful in this scenario.

Two parameters used by the algorithm are support and probability.Suppose there are two products in the shopping basket, number of instances with the combo of both the products in dataset give the support parameter. The part of instances for which dataset that has one product also contain the other is probability.

Creating mining models using PMML is not supported by Association algorithm. Use of OLAP mining models and creation of data mining dimensions are supported by this algorithm along with drill through.

5.Sequence Clustering Algorithm

As the name suggests in a Sequence Clustering Algorithm events in the given data is linked by following sequences. E.g. while browsing a website the order of clicks give a sequence. A customer adding products to a shopping cart also gives a sequence. By using this algorithm the company can find groups of similar users who browse through the website with similar click paths or can predict sale of a particular item.

Let take an example of an online shopping company which wants to display possible user interested products at the 1st page in its website's UI with its main objective to reduce the numbers of clicks a user does in the website to the purchase a product of user interest. In order for this the information about how the user clicks through pages and what all products/pages user is interested is needed to know customer can make use of Sequence Clustering algorithm. Once this information is known the customer can develop the website in such a way that the users can purchase products which they are interested without many clicks in 1st page itself. This in turn could effect in increased product purchase by customers and hence make more profit.

Sequence Clustering algorithm is actually a combination of clustering Algorithm with Markov chain analysis which identifies different clusters and their sequences. Sequence data which this algorithm used represents a series of transitions or events between states in a dataset. All the different transition probabilities and differences are measured by the algorithm. Distances between all the feasible sequences in the dataset are also measured. This is used to determine sequences that are the best candidates to be used as inputs for clustering. Once these candidate sequences are created, the sequence information is used as input in order for EM method of clustering.

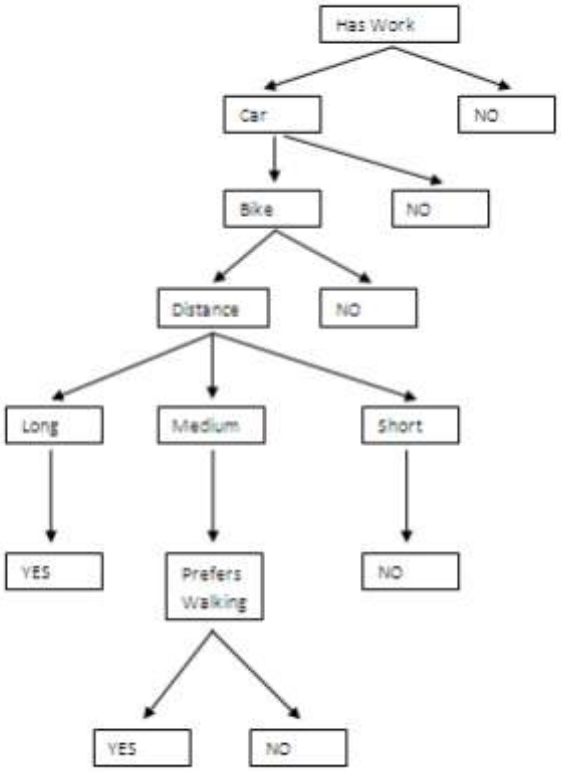
Drill through, using of OLAP mining models and creatingdata mining dimensions and creating mining models by using PMML are supported by this algorithm.

6.Decision Tree Algorithm

Suppose we have got a set of data containing several types or classes.

E.g. take the case of a bus service company and what they know is whether a customer will buy a ticket in the future. Here 'whether or not customer buys ticket' becomes the class. So here we can see a split - whether an unknown customer will buy something or not. So the idea is to ask questions. A series of questions' answer from the root (of the tree) will lead to a leaf node. This node provides the classification.

Given below decision tree representation for finding out whether customer will use bus in order for transportation to his/her workplace.



CONCLUSION:

Data mining is the integration of various techniques from multiple disciplines such as statistics, machine learning, pattern recognition, neural networks, and image processing and database management systems

REFERENCES:

1.http://msdn.microsoft.com/en-us/library/bb522495.aspx
2.http://en.wikipedia.org/wiki/Cluster_analysis
3.http://technet.microsoft.com/en-us/library/bb677216.aspx
4.Delivering Business Intelligence with Microsoft SQL Server 2008 – By Brian Larson
5.http://en.wikipedia.org/wiki/Time_series

Publish Research Article
International Level Multidisciplinary Research Journal
For All Subjects

Dear Sir/Mam,

We invite unpublished Research Paper,Summary of Research Project,Theses,Books and Book Review for publication,you will be pleased to know that our journals are

Associated and Indexed,India

- International Scientific Journal Consortium
- OPEN J-GATE

Associated and Indexed,USA

- EBSCO
- Index Copernicus
- Publication Index
- Academic Journal Database
- Contemporary Research Index
- Academic Paper Databse
- Digital Journals Database
- Current Index to Scholarly Journals
- Elite Scientific Journal Archive
- Directory Of Academic Resources
- Scholar Journal Index
- Recent Science Index
- Scientific Resources Database
- Directory Of Research Journal Indexing

Golden Research Thoughts
258/34 Raviwar Peth Solapur-413005,Maharashtra
Contact-9595359435
E-Mail-ayisrj@yahoo.in/ayisrj2011@gmail.com
Website : www.aygrt.isrj.org