ISSN: 2319-8753

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 1, January 2014

Modelling Structures in Data Mining Techniques

Ananth Y N [1], Narahari.N.S [2]

Associate Professor, Dept of Computer Science, School of Graduate Studies- JainUniversity- J.C.Road, Bangalore, INDIA [1]

Professor and Head, Dept of Industrial Engineering & Management, R.V.College of Engineering Autonomous-VTU – R.V.Vidyaniketan Post,Mysore Road,Bangalore-INDIA[2]

Abstract: Data mining involves finding out patterns of data from within large data sets-The large sets of data can be structured or unstructured-The data mining process involves two phases In the first step we develop data structures which can be used to hold the underlying data sets in a suitable manner-The second phase makes use of several algorithms to generate patterns or learn about the data. This process involves a data mining model -The data mining model encompasses the set of data, statistics and patterns that can be applied to new data to generate predictions and make inferences about relationships. This paper discusses some of the aspects related to how these structures can be created and manipulated for a variety of data. It also discusses how models can be created using several data mining techniques.

Keywords: Data Mining, Data Modelling, Big data, structured data, unstructured data, document databases, performance of students.

I. INTRODUCTION

The process of data mining has gained importance as the need to analyse big data is increasing. It is important to consider the various formats of this "big data" because the data structures required for modelling and analysing the data would depend on that. Basically, the different "types" of data sets include traditional SQL databases, raw text data, key-value stores, and document databases. While traditional SQL databases offer a highly structured means of data storage, they also put rigid constraints on the data that could be stored. Document databases are a totally new breed of data formats in which heterogeneous data could be stored and analysed. This paper discusses different kinds of data structures that could be used for modelling the data and also how models could be developed using those structures. Also, a particular application of the data mining techniques, namely in analysing the performance of students in higher education, is presented.

II. LITERATURE SURVEY

Survey of literature for this paper is done with respect to two aspects. The first one is to study the different data mining algorithms available. Some of the common data mining algorithms are in the categories of clustering, classification and so on. Data mining algorithms such as k- means, are powerful algorithms which could be used with a variety of kinds of data. A host of algorithms are available in the literature which finds their application in variety of situations across many fields such as business, education, and so on. The second part of the literature survey includes studying the actual implementation aspect of these algorithms – where in we study about different structures of data that could be used. Structures to store data as well as retrieve and analyse data are found exhaustively in the literature. Traditional structures like text files, sql databases have been found and used in applications- and recently – document databases, which are useful in modelling web pages, are also being used to analyse complex patterns of data spread across many sources.

ISSN: 2319-8753

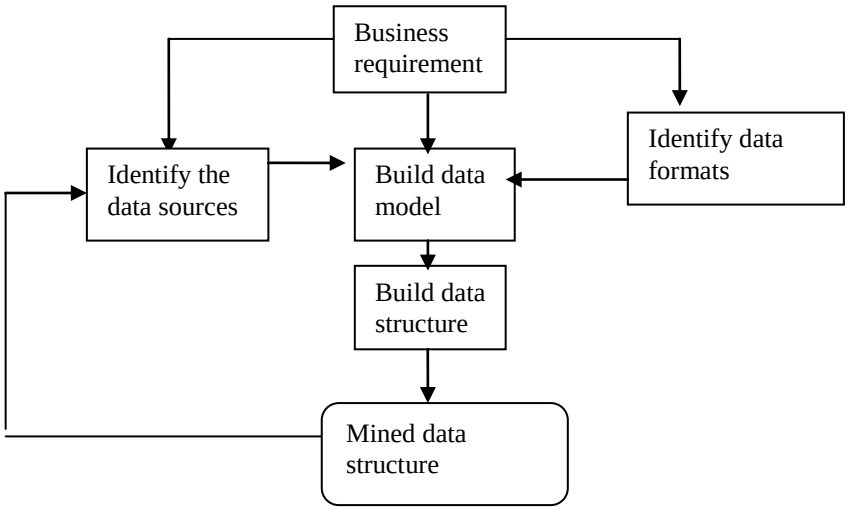
International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 1, January 2014

III. PROCESSES IN BUILDING STRUCTURES

It is important to understand the following sequence of processes in building the data structure and modeling. The following diagram would give a picture of this sequence.



IV. TEXT MINING

Text mining, involves deriving high quality information from the text. This information is typically derived from methods such as statistical pattern learning applied to large sets of text data. Proper structuring of the text data would involve parsing of the textual data, addition/deletion of some linguistic features and insertion into database objects. This structured data would be used to derive patterns of data and finally evaluation and interpretation of the data involved.

The goals of Text analysis would be information retrieval, lexical analysis to study word and frequency distribution, pattern recognition, data mining techniques including link and association analysis and predictive analytics. The main focus is to turn text into data for analysis by applying Natural Language processing and analytical methods.

V.DATA MODELS AND STRUCTURES W.R.T RELATIONAL DATA

Data mining models are a set of data columns along with the data mining algorithms that are used for data mining. When a certain data is mined there are two steps – running the mining algorithms on a set of training data-The data mining model content created by the training process is stored as the data mining model content.
In this scenario, it is important to distinguish between data mining models and data mining structure. The data mining structure contains the information that defines the data source. Whereas the data mining model stores information derived from the statistical processing of the data such as patterns and relationships.
A mining model is empty until the data is analysed by using the algorithms. The mining model contains the results of the mining process- along with the metadata and bindings back to the mining structures.
The metadata consists of the name of the model, definition of the model, including the specific column names that were used in the mining process, the definitions of any filters that were used while processing the data, and the algorithm that was used to process the data.
The model depends to a large extent on what is being analysed and the choice of the algorithms used to analyse. For example the same data set could be analysed with a clustering algorithm, a decision tree algorithm or a naïve Bayes algorithm. Each of these give out their own set of patterns, itemsets , rules or formulas, which can be used for making predictions. Each of the algorithms behaves in their own way and hence the “content “of the model also is organized in

ISSN: 2319-8753

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 1, January 2014

different kind of structures. In some of the models the data and the content might be organized in the form of clusters and in some others it might be in trees.
The model is also affected by the data that is used to train it on. That is, the same set of algorithms on the same set of architectures could yield different results if different training data set or different filters is used. But the model itself is not used to store the data. It is used to store the results of the analysis like the summary statistics, patterns and such. The mining structure consists of the actual data.

The general steps that could be used in creating a data mining model would be the following.

- Create the underlying mining data structure and include the columns of data that might be included.
- Select the best suited algorithm for the analytical task.
- Choose the particular columns from the structure that could be used in the modeling,and specify in what way they have to be used
- Optionally, set the parameters to fine tune the processing by the algorithm

The following points have to be emphasized here. Even though a large amount of data gets generated as a result of the training process, the training data itself will not be stored. Only the analysis data generated out of the processing of the training data and the column definitions used to generate the data gets stored as the content of the model.

VI.DATA MINING MODELLING STRUCTURES

The mining structures define the data from which the mining models are built. It specifies the source data view, the number and type of the columns, and an optional partition into a training set and a test set of data. A single mining structure can support a number of mining models that share the same domain.
Setting up a mining structure involves the following steps.

- Define a data source[1]
- Select columns of data to include in the data structure and defining a key
- Define a key for the structure, including a key for the nested structure , if applicable
- Specify whether the data should be belonging to a training set or a testing set
- Process the structure

The data sources for mining structures could be the set of columns that are available in the existing data source. The properties of these set of columns could be used, to modify the data types, create aggregations or alias columns. Etc.
If multiple models are built from the same data source, then different set of columns could be used for it. For example, a single data set could be used where in different columns set could be used to define a decision tree algorithm or a clustering algorithm.

VII.DATA MINING MODEL NODES

The content of a data mining model usually consists of the mining nodes. Each node contains the information about the attributes needed to define the node, the relevant rules to process a case against the node, and the analysis gained from training the node. Each node can also be related to other nodes to further support the complexities of the algorithms, for example – decision trees and clustering algorithms in a common structure. The data mining nodes can be further browsed to understand the decisions or aggregations done by the algorithm involved and they can be modified to further adjust the model. The following six types of nodes can be recognized as of now.

Model:

A model node is the topmost node in any data mining model, regardless of the actual structure of the model. All models start with a model node.

ISSN: 2319-8753

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 1, January 2014

Tree

For all tree based nodes this node serves as the root node of the tree. A data mining trees might have many trees that make up the whole, but there is only one tree node from which all the other nodes are related for each tree.A decision tree model has got one model node and at least one tree node.

Interior

An interior node represents a generic node of a model. For example in a decision tree, this node usually denotes a split in the tree.

Distribution

A distribution node is guaranteed to have a valid link to a nested distribution table. A distribution node describes the distribution of values for one or more attributes according to the data represented by this node. A good example of a distribution node is the leaf node of a decision tree.

Cluster

A cluster node stores the attributes and data for the abstraction of a specific cluster. In other words, it stores the set of distributions that constitute a cluster of cases for the data mining model. A clustering based model has one model node and at least one cluster node.

Unknown

The unknown node type is used when node does not fit any of the other node types provided and the algorithm does cannot resolve the node type.

VIII.DATA MINING MODEL PROPERTIES

With certain data mining environments, a data model essentially has certain properties. They usually involve the name of the model, description, the date the model was last processed, permissions on the model, and any filters on the data that were used for training.

Each mining model can also have properties that are derived from the mining structure, and that describe the columns of data used by the model. If any column used by the model is a nested table, the column can also have a separate filter applied.

The following two properties can be considered as important.

1. **Algorithm property**

Specifies the algorithm used to create the model. The **Algorithm** property applies to the model and can be only one time for each model. A change in this property might make some of the columns to be invalid –A reprocessing of the model is always needed once this property is changed.

2. Usage property:

Defines how each column is used by the model. The column usage can be set to Input, Predict, Predict only or key. The usage property applies to individual mining model columns and must be set for every individual column that is included in the model. If the structure consists of a column that is not used in the model, then it is set to Ignore. Example of a column that can be present in the structure but not considered in the analysis is a customer email id column**.**

IX. MINING MODEL COLUMNS:

A mining model contains the columns that are derived from the columns defined in the mining structure .The columns to be included in the mining model can be chosen from the structure and copies of the columns can be created and the columns renamed. As part of the model building process, the usage of each of the columns can also be set.

Care has to be taken while choosing a column to be part of the model, because some of the columns might represent the same data .Some columns having unique values have to be excluded. Provisions are available to flag a particular column if it is not required for the analysis to ignore that in the model. This means that the column is present in the structure but is not part of the model. The data could be retrieved later using drilldown features of the environment.

Depending on which column is chosen for mining, certain columns might be useful for certain analysis and certain columns might not. For example if a data source contains a numeric attribute like marks of the student and the model requires discrete values, the data might have to be converted to discrete ranges or it might have to be removed. In some

ISSN: 2319-8753

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 1, January 2014

cases the algorithm automatically bins the data but the results might not be as expected. In such cases additional copies of the column can be made and different models could be tried.
The mining model can be made suited for the particular use by doing some of the following.

- Use different columns of data in the model, or change the usage, content type, or discretization method for the columns.
- Create filters on the mining model to restrict the data used in training the model.
- Change the algorithm that was used to analyse the data.
- Set algorithm parameters to control thresholds, tree splits, and other important considerations.

X.MODELLING DOCUMENT DATABASES:

Traditional RDBMS environment is widely used in Business Intelligence and other areas to store the data. These follow a strict structure and require significant amount of preparation with regard to the schema. The document based databases make this quite easy because information can be dumped in a flexible format. Additionally, new methods of analysing the fixed data could be found out.

Document database Architecture:

One of the key elements of document databases is that they can work with much larger structures and datasets than normal. In particular, because of their distributed nature and the different way in which they store the data physically, they are ideal when a vast amount of data has to be processed, as is often the case with data mining. The following are some of the important features of document databases.

Schema Less: Document databases do not have a pre defined structure of their own. Unlike RDBMS ,where in the structure of a database is specified with definitions of the tables , with a document database, the information can be stored into the documents without having to worry about the structure, whether there are multiple fields and even in most cases , the one to one and one to many relationships. Instead one can concentrate on the content of the database itself. More flexibility in processing the data is also there- for example to collate data from Twitter, face book and other social networking sites searching for patterns.

Logical structure: In traditional RDBMS data, data is analysed using individual tables and columns within the tables. In order to integrate some data, we have to collate data from different sources and different perspectives. Document data lessens this kind of a work by treating a piece of data as a logical unit – for example one can look at a web page and take individual components out of it and treat them as belonging to the same object.

Migratory structure: In any data related application, data changes over time and this affects the applications that are using the data. This means that because the underlying data structure is fixed, adapting to newer formats of the original data is difficult and complex process. But with a document based format, the whole document is treated as an unit and hence the data structure can be modified.

A popular format for storing document data is JSON, an object notation format from the Javascript language. It allows us to store string, numbers, arrays and record (hash) data and combinations of those types. With document based structures, and technologies like Hadoop, it is the "processing at extraction "that makes these to be powerful.

In a typical RDBMS, the table structure is known and to get a particular result, we can issue a query depending on the table design. The process of the information is done at the point of input, separating out the information so that it can be inserted into tables and then unified at the point of output by recombining the information. This demands that the table design is known and an SQL query can be written only when the structure of the table is known.

With a document based format, it is the process of the raw data that creates the harmonized view of the information that enables the data to be processed, whether that is value based or thematic or textual. The information is put into multiple documents, and map/reduce system processes that information and generates a structure table from this data.

ISSN: 2319-8753

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 1, January 2014

XI. MODELLING STRUCTURES FOR DATA MINING IN ANALYZING PERFORMANCE OF STUDENTS IN HIGHER EDUCATION

A particular application of data mining techniques, which is quite useful, is in analysing the performance of students in examinations like PGCET examinations. Here, there are distributions of marks available in large data sets and they could be analysed to find out patterns and clusters. The utility of this analysis is that they could be used for understanding the students' ability.

The basic assumption made here is that students' performance is progressive, which means students performance level gets reflected in their marks – and they are constant for long years of their education period. This leads to the assumption that a student's performance at the secondary school level to the college level to the undergraduate level – could be used to indicate his/her level of performance in PG level and hence, to a large extent, they show up in their performance in entrance exams like PGCET. Based upon this assumption the scores of the students in the PGCET has been analysed and compared with their marks at the degree level. The comparisons are mainly between the marks of a student in the graduate level and the entrance exam on an year-on –year basis and after that on a subject wise basis. For example, the marks of students who are graduating in the year 2013 and taking up the PGCET are analysed with the marks of the students graduating in the years 2011 and 2010.The correlation which could be found is that between the PGCET marks of the three years 2010, 11 and 2012, we can see certain definite patterns. There could be some distribution of the marks which shows that in a particular subject, say Computer Science, there are clusters of students who score the same marks for obtaining the same ranks. For example ranks 100 to 200 go in the same clusters although the exact marks scored to secure those ranks might be different. The second type of analysis involves comparing the marks at the graduate level and the marks in the entrance exam. This analysis gives clusters which indicate whether there is a direct relationship of the marks at the undergraduate level and the entrance exam marks. A set of clusters indicate a close relationship between the two sets of marks. With these clusters we can definitely say that high achievers at the graduate level remain high achievers at the PG level. Some other clusters indicate not so close a relationship between the two sets. These clusters indicate that students performance at the graduate level either gets to be better by the time they take up the entrance or they may be get worse. It is possible from data mining techniques to extract these kinds of patterns. The main techniques which could be used here are clustering and classification. We could use clustering techniques such as k- means to generate the clusters at a gross level , then refine them with definite values of the starting mean values at the subsequent stages. Once the clusters are identified, we can use them to determine what kind of a cluster each one is- here the correlation techniques become important. Clusters with the same relationship between the sets of marks, or clusters with different relationship could be identified. From this the students performance could be assessed.

The basic structures which could be used here as the input structures to the data mining process could either be containing fundamental data types such as arrays and structures or flat files. But for analysis and correlation of a greater number of fields , for example analysis across the dimensions of year, subject, qualifying exam marks, entrance exam marks – and depicting them in a user friendly manner, the arrays have to be multidimensional arrays- each of the dimensions storing one type of marks-and for visual depiction – we have to use multidimensional graphs- which could have , for example , a graph with 4 coordinate axes-or a 3D-graph which is rotatable – to show the additional dimensions whenever the user chooses to see them.

A whole model of analysis could be built out of this kind of a study which could be used to predict, at least approximately as to what might be the performance of students in the entrance exams, given the past performance and performance at the graduate level. The predictive model would include sets of correlated marks; all deduced using the different algorithmic techniques described here. To create and analyse more number of clusters and also when the qualifying marks is not available, we can simulate the performance using random numbers and what if analysis techniques. To do this, initially we build the correlation from the available data. Then we generate random numbers starting with a particular seed value which would generate the same distributions-for example we could use seed values to generate the marks distribution where the correlation between the two sets of marks is a certain value , say 0.6900.We can do repeat this to all the available sets of data and then strike a relationship between the seed value and the distribution. Further, this could be used generate sets of relationships which would include all kinds of possible

ISSN: 2319-8753

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 1, January 2014

relationships and hence to predict what might be the performance in the entrance examination, given a set of qualifying marks. To achieve this , at the first round of the analysis, data available could be used and for the further levels where in we have to consider different seed values as inputs, we could use arrays to store random numbers and feed them to graph generators. Different programming languages implement this kind of a storage using different structures. C++ uses arrays and java uses arraylists to do this.

XII.CONCLUSIONS

This paper has discussed some of the popular structures available for use when it comes to data mining and analysis of "big data". Although "big data" is a new buzzword, techniques related to analysis of data in huge volumes is becoming more of a necessity. The structures described here are self containing but are not exhaustive. Techniques like Map reduce will be important ones to study and adopt in the future. This is an attempt towards understanding the different formats of structures wrt the traditional data formats as well as document databases. Also discussed is a particular case study, that of analysing the students' performance, wherein data mining techniques could be used.

.

REFERENCES

[1] www.google.com