ISSN2394-3777 (Print)
ISSN2394-3785 (Online)
Available online atwww.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 5, Special Issue 10, March 2018

Review Paper on Data Mining: Techniques and Architecture

Rajdeep Kaur

Assistant Professor, P.G. Department of Computer Science and Application, S.G.G. Janta Girls College, Raikot, Punjab, India

Email: kaurrajdeep758@gmail.com

Abstract :- Basic data mining techniques and architecture is introduced by this paper. Data Mining is a process to find the helpful information from database, the data in the database may be incomplete, noisy or random data. Various techniques of data mining are here to extract the data from databases. These techniques are (regression, classification, prediction, clustering, decision trees, association etc.) used by data mining systems. Various methods like Neural Networks, Pattern Recognition, Machine Learning, Knowledge Based Systems, Statistics, Visualization Methods etc. are used by data mining process. Various data mining algorithms are applied on the data to find the better results. Pruning techniques are used to remove the noisy data from the database. These techniques are also used to reduce missing values and incomplete data. Four types of architecture are here that are No-Coupling, Loose-Coupling, Semi-Tight Coupling and Tight-Coupling. Tight-Coupling has three tiers :- Data Layer, Data Mining Application Layer and Front-End Layer. Tight-Coupling architecture provides high scalability, good performance and integrated information.

Keywords: - data mining, classification, decision trees, pruning, tight coupling.

I. INTRODUCTION

Data Mining is the technique to discover useful information from database. It has different methods to arrange and collaborate large data including Classification, Clustering, Association and Sequential Patterns. Machine Learning, Neural Networks, Visualization Methods and Statistical Analysis also used by the data mining process. It is also use the biometric techniques to recognize different number of patterns without having any knowledge about the data.

Data Mining perform various tasks. These tasks are divided into two categories – descriptive and predictive. General properties of the database are characterized by the descriptive tasks. Predictive mining tasks are used to predict the future data on the basis of current data.

II. TECHNIQUES OF DATA MINING

A. Association

Association is a technique of data mining to recognize any pattern on the basis of the relation of items that are using in same transactions. This technique is also known as Relation Technique. This technique is used by Market Basket Analysis to find the frequently purchased items by the number of customers. Retailers, to increase their sales, use these Market Basket Analysis techniques. Two parts of association are if and then.

If/Then patterns are used to analyzing data, to create Association Rules. Support and Confidence criteria is used by these Association rules. Support indicates, how frequently data items are appear and confidence indicates number of times the statements found to be true. Association plays an important role in Product Clustering and Catalog Design. Association rules are also defined by the programmers to develop a Machine Learning program.

B. Clustering

Clustering is a technique that is used to divide the data into valid or purposeful sub classes, are known as Clusters. These Clusters are created according to the different kind of information. Similar information is contained by one cluster and dissimilar by another one. Hard clustering and soft clustering are used in clustering.

In hard clustering, only single cluster contains the same object.

In soft clustering, more than one cluster contains the same object.

Clustering is used in Pattern Reorganization, Image Processing and in Market Research. With the help of clustering different animal taxonomies, genes with similar functionalities, are indentified in the field of Biology. Group of houses are also identify according to the values, location and type of the house.

ISSN2394-3777 (Print)
ISSN2394-3785 (Online)
Available online atwww.ijartet.com
International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 5, Special Issue 10, March 2018

Clustering is also used to grouping/classifying documents on the web. It also helps to detect credit card frauds by using in detection applications/algorithms. Cluster analysis is a tool that is used to define/find characteristics of each cluster.

- *Requirements of Clustering*
    1. For large databases we need scalable clustering algorithms .
    2. These algorithms should be able to deal with any data such as numerical, binary etc.
    3. Both low dimensional and high dimensional data should be handle.
    4. Missing, Erroneous and Noisy data is also contained by the databases, so sometimes quality of clusters is not good.
    5. Clusters with arbitrary shape should be detected by these algorithms.

Different types of methods are used by the clustering such as Partitioning Method, Hierarchical Method , Density-Based Method, Model-Based Method, Constraint-Based Method and Grid-Based Method. K-means Clustering algorithm is used for clustering precise data and uncertain data.

C. Classification

Classification is a technique that is based on machine learning. Classification technique is used to classify the data from the database into different predetermined classes or groups. This method uses different mathematical techniques/methods like Neural Networks, Statistics and Decision Trees. Classification technique has two steps. In first step, data mining classification algorithms are applying on the data, and in second step data is grouped/classified according to the predetermined classes and groups. First step is Model construction and second is Model usage. In Model Construction, we define the set of tuples, labeled data, classification model that are represented by classification rules/decision trees or mathematical formulae. In Model usage, we test/measure accuracy of the model, after testing if model is accurate then we use the model.

Classification Models :-

- Classification using decision tree induction
- Neural Networks
- SVM (Support Vector Machines)
- Bayesian Classifications
- Rule Based Classification using if-then rules

D. Prediction

This technique is used to find the relationship between dependent or independent variables and between independent variables. For example, by using this technique, we predict the profit for the future on the basis of sale, so here, sale is an independent variable and profit could be a dependent variable. Basically, Prediction finds the relationship between a variable you know and a variable you need to find for future. In credit card fraud detection, we first analyze the card usage by the person, if we detect any abnormal pattern , it should be reported.

According to the available data/basis on the available/present data, we can predict the future values, here two types of predictions are :-

- Deterministic Predictions
- Probabilistic Predictions.

Neural Networks, Bayesian Classifiers, Temporal Belief Networks, Nearest Neighbor etc. are some techniques of Prediction.

Regression techniques is also used for prediction. Using

Regression technique we find the relationships between variables.

1. Various regression methods :-

- Linear Regression
- Nonlinear Regression
- Variable Linear Regression
- Variable Nonlinear Regression

2. Classification and Prediction Issues

a) *Data Mining:-* Techniques are used to reduce the noisy and missing data.
b) *Relevance Analysis:-* Eliminate irrelevant data.
c) *Reduction and Data Transformation* **:-** Generalization of Data and Normalize the data in the database.

E. Sequential Patterns

Sequential Pattern method/technique is used to recognize/find similar patterns or trends over a time period. This technique is used to create well-organized databases, catalogue (index) sequential information, recognize frequently patterns, compare sequences, to find similarities and recover missing data from sequences. Two types of sequence mining are string mining and itemset mining. String Processing Algorithms are used for String Mining and association rules are used for Itemset Mining. More complex patterns such as choices, loops are extended from local process models of Sequential Mining.

ISSN2394-3777 (Print)
ISSN2394-3785 (Online)
Available online atwww.ijartet.com
International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 5, Special Issue 10, March 2018

1. *String Mining :-* String Mining uses limited alphabets to represent items in a sequence, an arrangement of the alphabets is used to determine properties of Gene and Protein by examine their sequence, so string mining is used by the biology application.
2. *Itemset Mining :-* Itemset Mining is used to find the frequent items and their sequence to find regularities of frequent occurring items in transactions, so itemset mining is by the Marketing Applications.

Sequential Patterns commonly use GSP Algorithms, Prefix Span, SPADE (Sequential Pattern Discovering using Equivalence classes).

F. Decision Trees

It is an easy to understand technique of data mining for users. A simple question is the root of the decision tree that has more than one answer. Each answer has its own further questions to determine the data to take the final decisions. For example :- Decision Tree to find whether or not to play Tennis.

In figure 1, starting node is the Outlook that has many answers i.e. Sunny, Rainy, Overcast. If overcast then we can play. If Rainy then we further check the conditions, if wind is weak then we can play the tennis otherwise we cannot. If sunny then we check the condition on humidity, if humidity is normal then we can play.

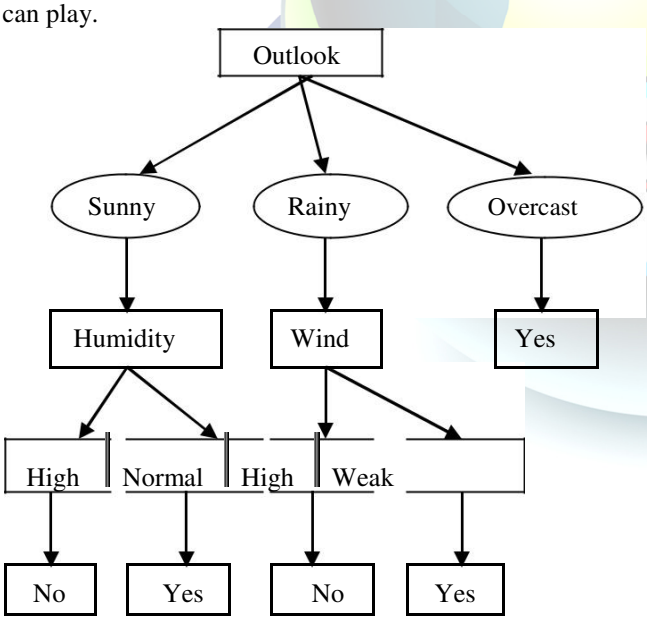

Fig. 1. Example of an Decision Tree

In 1980, J. Ross Quinlan, a machine researcher developed a algorithm ID3 (Iterative Dichotomiser), a decision tree algorithm. Later C4.5 algorithm (successor of ID3) presented by him. Greedy approach is used by these algorithms. Divide and Conquer technique is used to construct decision trees in these algorithms, no backtracking is used.

1. Pruning Technique of Decision Trees (Tree Pruning)

In database, may be some missing values or incomplete data or noisy data exist, so to eliminate or to reduce this type of data tree pruning is used. Decision tree algorithms are not able to find the noisy or inaccurate data, so this technique is used for this. By using tree pruning technique we can make the trees simple, less complex, smaller and easy to understand.

Two Pruning approaches are here :-

1. Pre-Pruning:- Pruning technique is applied on data before we construct the decision tree for that data.
2. Post-Pruning :- Pruning technique is applied on data after we construct the tree properly. It is a two stage method.

III. ARCHITECTURE OF DATA MINING

Now days, Databases and Warehouses are used to store the data so question is, should we design a system such as data mining system that is used to coupling and decoupling with database or warehouse.

So, here four possible architectures of Data Mining Systems are exist :-

A. No-Coupling

Where data mining system does not use any function of the database or warehouse system, that architecture is known as no-coupling architecture. This system access the data from the source file, process it by applying the data mining algorithms and stores results in another file. This architecture does not utilize any advantage of the database such as already efficient in organizing, storing, retrieving. This architecture considered as poor architecture, so it is used only for simple processes.

B. Loose-Coupling

When the database and data warehouses are used to access the data, that architecture is known as loose-coupling architecture. In this architecture data mining system accesses the data from database, process the data by applying data mining algorithms

ISSN2394-3777 (Print)
ISSN2394-3785 (Online)
Available online atwww.ijartet.com
International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 5, Special Issue 10, March 2018

and stores result into those systems. Loose-Coupling architecture is used by those systems where high scalability and performance does not require i.e. Memory based data mining systems.

C. Semi-Tight Coupling

in this architecture, we links the systems to the database, so data mining systems utilizes some functionalities of the databases to perform various tasks. These tasks are indexing, aggregation, sorting etc. Intermediate results can also stored in database to enhance the performance.

D. Tight-Coupling

In this architecture, Databases are used as an component of data mining system to access the information using integration. All functionalities of database are utilized by the data mining systems to perform various tasks of data mining. This architecture is used by those systems where high scalability, integrated information and good performance is required.

This architecture has three tiers :-

1. Data Layer :- Databases or warehouses are in this layer. This layer provides an interface to data sources. Results after applying data mining algorithms are also stored in this layer, so this layer is also used to present results to end-users in form of reports or any other form.
2. Data Mining Application :- To access the data from database, data mining application layer is used. Here various transformation techniques are used to transform the data in given format and then we process the data by applying the data mining algorithms.
3. Front-End Layer :- It provides a friendly interface to end-user. By using this interface end-user can interact with the system. Results are presented in the visualization form such as reports to the end-users in this layer.

IV. CHALLENGES OF DATA MINING

- Scalability
- Heterogeneous Data
- Complex Data
- Quality of Data
- Ownership of Data
- Distribution of Data
- Dimensionality
- Privacy Preservation
- Streaming Data

V. CONCLUSION

Data Mining is not only a technique to extract data, it is also used to find the relationships between the dependent or independent variables, to divide the data into different number of clusters where each cluster contains similar data. Data Mining process is also used two classify the data according to their functionalities and by using this technique we also predict the future values basis on the present values. These techniques are helpful in various fields i.e. agriculture, biology etc. Various data mining architectures are available according to the data and need of the users. No-Coupling architecture is used to process simple data because this architecture is not scalable. Tight-Coupling is used where high scalability is required. This architecture also provides integrated information. Pruning Techniques are available in decision tree technique to reduce incomplete and noisy data. Data Mining extract the data , transform it and store results. It is also used to manage the multidimensional database. Data Mining systems provides the results to the user in understanding format.

REFERENCES

[1] Hemlata Sahu, Shalini Shrma and Seema Gondhalakar, A Brief Overview on Data Mining Survey, Vol. 1, Issue: 3, IJCTEE, ISSN 2249-6343.

[2] Qing-yun Dai, Chun-ping Zhang and Hao Wu, Research of Decision Tree Clasification Algorithm in Data Mining, Vol.9, pp. 1-8, 2016.

[3] Sagar S. Nikam, A Comparative Study of Classification Techniques in Data Mining Algorithms.

[4] Nikita Jain, and Vishal Srivastava, DATA MINING TECHNIQUES: A SURVEY PAPER, IJRET, Vol. 02,Issue: 11, Nov-2013, Available on http://www.ijret.org.

[5] Mrs. Bharati M. Ramageri, DATA MINING TECHNIQUES AND APPLICATIONS, Vol. 1,

[6] Mansi Gera and Shivani Goel, Data Mining- Techniques, Methods and Algorithms : A Review Paper on Tools and their Validity, Vol. 113, 2015.

[7] Yihao Li, DATA MINING: CONCEPTS, BACKGROUNDS AND METHODS OF INTEGRATING UNCERTAINTY IN DATA MINING.

[8] https://www.tutorialspoint.com/data_mining/dm_dti.htm

ISSN2394-3777 (Print)
ISSN2394-3785 (Online)
Available online atwww.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 5, Special Issue 10, March 2018

BIOGRAPHY

Author Name: Ms. Rajdeep Kaur
Designation: Assistant Professor
Department: Post Graduate Department of Computer Science and Applications
Qualification: B.C.A., M.C.A.