Survey on Data Mining Techniques, Process and Algorithms

K.K.Nivethithaa1st, Dr.S.Vijayalakshmi2nd
Ph.D Research Scholar1, Associate Professor2
Department of Computer Science
nivethithaakalimuthu@gmail.com1, vijics@srcw.ac.in2
Sri Ramakrishna College of Arts and Science for Women, Coimbatore, Tamil Nadu, India

Abstract. The term “Data Mining” is refers to the extraction of patterns and knowledge from large amounts of raw data and often defined as finding hidden information in a database. It insinuate analyzing data patterns in large volume of data using one or more software. Data mining involves effective data collection and warehousing as well as computer processing.

Keywords: Data Mining, Classification, Prediction, data mining tools, algorithm

1. Introduction

Data mining describes the concept of discovering knowledge from databases which includes the process of finding anomalies, patterns and correlations within large data sets to predict outcomes. Using a boundless techniques these information are used to increase revenues, cut costs, improve customer relationships, reduce risks.

2. Elements Of Data Mining

Data mining consists of five major elements:

- Extract and transform the data and load the transformed data onto the data warehouse system.
- Store the data and manage the data in a multidimensional database system.
- Provide data access to business analysts and information technology professionals.
- Analyze the data by application software.
- Present the data in a useful formats such as a graph, table etc.

3. Goals Of Data Mining

1. Prediction
- Determining how certain the attributes will behave in the future.
For example, how much sales volume a store will generate in a given period.

2. Identification
   - Identify patterns in data.
   - For example, newly wed couples tend to spend more money buying furniture’s.

3. Classification
   - Partition data into classes.
   - For example, customers can be classified into different categories with different behavior in shopping.

4. Optimization
   - Optimization of the limited resources such as time, space, money or materials.
   - For example, how to best use advertising to maximize profits (sales).

1.1 GOALS OF DATA MINING

4. Types of Data

Data mining is executed on following types of data

- Relational databases
- Data warehouses
- Advanced DB and information repositories
5. Data Mining Techniques

1. Classification:
- Classification analysis is used to retrieve important and relevant information about data, and metadata.
- This data mining method helps to classify data in different classes.

2. Clustering:
- Clustering analysis is one of data mining technique used to identify data that are like each other.
- This process helps to understand the differences and similarities between the data.

3. Regression:
- Regression analysis is the data mining method of identifying and analyzing the relationship between variables.
- Regression is used to identify the likelihood of a specific variable, given the presence of other variables.

4. Outer detection:
- Outer detection is a data mining technique which refers to the observation of data items in the dataset which do not match an expected pattern or expected behavior.
- Outer detection is used in a variety of domains such as intrusion, detection, fraud or fault detection etc.
- Outer detection is also known as Outlier Analysis or Outlier mining.
1.2 DATA MINING TECHNIQUES

5. Sequential Patterns:

- Sequential Patterns is data mining technique which helps to discover or identify similar patterns or trends in transaction data for certain period.

6. Prediction:

- Prediction has used a combination of the other techniques of data mining like trends, sequential patterns, clustering, classification, etc.
- It analyzes past events or instances in a right sequence for predicting a future event.

7. Association Rules:

- This data mining technique helps to find the association between two or more Items.
- Association Rules discovers a hidden pattern in the data set.

6. Areas of Data Mining Implementation

1. Healthcare
2. Education
3. Market basket analysis
4. Customer relationship management (CRM)
5. Manufacturing engineering
6. Finance and banking

7. Fraud detection
8. Monitoring Patterns
9. Classification
10. Association
11. Anomaly Detection
12. Clustering
13. Regression
14. Prediction
15. Sequential Patterns
16. Decision Trees
17. Visualization
18. Neural Networks
19. Data Warehousing

7. Data Mining Process

Before the actual data mining could occur, there are several processes involved in data mining implementation and they are listed in the following steps.

**Step 1: Business Research** – Before the commencement, one need to have a complete understanding of the enterprise’s objectives, available resources, and current scenarios in alignment with its requirements.

**Step 2: Data Quality Checks** – Data gets collected from various sources, so it needs to be checked and matched to ensure no bottlenecks in the data integration process.

**Step 3: Data Cleaning** – Data cleaning involves selecting, cleaning, formatting, and anonymizing data before mining.

**Step 4: Data Transformation** – It compresses of five stages, Data Smoothing, Data Summary, Data Generalization, Data Normalization, Data Attribute Construction.

**Step 5: Data Modeling:** For better identification of data patterns, several mathematical models are implemented in the dataset, based on several conditions.
### 1.3 DATA MINING ALGORITHMS

| ALGORITHM          | TECHNIQUE             | ADVANTAGE                                      | DISADVANTAGE                                                                 | SUPERVISED OR UNSUPERVISED | TOOLS IMPLEMENT THE ALGORITHM |
|--------------------|-----------------------|-----------------------------------------------|-------------------------------------------------------------------------------|----------------------------|--------------------------------|
| **C4.5**           | Constructs a classifier in the form of a decision tree | Builds models that can be easily interpreted, easy to implement and deals with noise | Small variation in data can lead to different decision trees. Does not work very well on a small training set | **SUPERVISED**             | OpenTox Orange |
| **K-means**        | k-means creates $k$ groups from a set of objects so that the members of a group are more similar | computationally faster than hierarchical clustering, produce tighter clusters | Difficult to predict K-Value. | **SEMI-SUPERVISED** | Apache Mahout, Julia, R, SciPy, Weka, MATLAB, SAS |
| **SVM (Support vector machines)** | SVM performs a similar task like C4.5 except SVM doesn’t use decision trees at all. | Effective in high dimensional space, memory efficient | Not suitable for large data sets, not perform very well when data set has more noise | **SUPERVISED** | scikit-learn, MATLAB, libsvm |
| **Apriori**        | The Apriori algorithm learns association rules and is applied to a database containing a large number of transactions | Easy to implement, use large itemset property | Requires many database scans, very slow | **UNSUPERVISED** | ARtool, Weka, Orange |
| **Expectation-Maximization (EM)** | the EM algorithm iterates and optimizes the likelihood of seeing observed data while estimating the parameters of a statistical model with unobserved variables | likelihood will increase with each iteration. | It has slow convergence | **UNSUPERVISED** | Weka, R, Scikit-learn |
| **PageRank**       | PageRank is a link analysis | PageRank is a global measure | It favours the older pages | **UNSUPERVISED** | Google |
| Algorithm       | Description                                                                 | Advantages                                                                 | Disadvantages                                                                 | Library/Platform                                                                 |
|-----------------|------------------------------------------------------------------------------|----------------------------------------------------------------------------|--------------------------------------------------------------------------------|---------------------------------------------------------------------------------|
| Algorithm       | Designed to determine the relative importance of some object linked within a network of objects. and is query independent |                                                                            |                                                                                 |                                                                                 |
| AdaBoost        | AdaBoost is a boosting algorithm which constructs a classifier.               | Used to improve the accuracy of weak classifier, it is flexible            | Extremely sensitive to noise data and outliers                                   | Supervised: Scikit-learn, ICSIBOost, Gbm Generalised boosted regression model |
| kNN             | kNN just stores the labeled training data.                                   | Quick calculation time, Simple algorithm, Versatile, High accuracy         | Prediction stage might be slow, Require high memory, computationally expensive   | Supervised: MATLAB k-nearest neighbour classification, Scikit-learn, k-nearest neighbour classifier k-nearest neighbour classification in R |
| Naive Bayes     | Naive Bayes is not a single algorithm, but a family of classification algorithms that share one common assumption | Works quickly suited for categorical input variables                      | Estimations can be wrong in some cases                                          | Supervised: Orange, Scikit-learn, R, Weka                                       |
| CART (classification and regression trees) | It is a decision tree learning technique that outputs either classification or regression trees | Fastest algorithm, Requires minimal supervision | Limited number of positions to accommodate available predictors                  | Supervised: Scikit-learn, R, Weka, MATLAB                                         |

9. Conclusion
Data mining plays a major role in all the fields in the future. Data mining has the power to transform a data into useful information with the help of tools which result in growth of technology in all the sectors. Artificial Intelligence and business learning etc new concepts are giving their effective performs with the support of Data mining.

References

[1].https://www.topicsforseminar.com/2019/06/data-mining-seminar-topics.html?m=1

[2].https://www.engpaper.com/data-mining-2019.htm

[3].https://www.researchgate.net/publication/332233710_DATA_MINING_CLASSIFICATION_TECHNIQUES_ON_THE_ANALYSIS_OF_STUDENT%27S_PERFORMANCE

[4].https://www.semanticscholar.org/paper/Application-of-Data-Mining-%E2%80%93-A-Survey-Paper-Sharma-Sharma/77c10addbcd9bd9e0ba1c00ccd3bb83f43c21c09

[5].https://arxiv.org/ftp/arxiv/papers/1211/1211.5723.pdf

[6].https://www.ripublication.com/ijaer18/ijaerv13n7_112.pdf

[7].https://www.talend.com/resources/what-is-data-mining/

[8].https://medium.com/rs21/ai-in-government-part-two-data-mining-502b2d40374d

[9].https://www.careerride.com/question-10-Data-Mining-1

[10].https://www.iedunote.com/data-mining

[11].http://people.cs.pitt.edu/~chang/156/21mining.html

[12].https://onix-systems.com/blog/8-data-mining-techniques-you-must-learn-to-succeed-in-business

[13].https://www.guru99.com/data-mining-tutorial.html

[14].https://www.upgrad.com/blog/data-mining-techniques/
[15].http://www.computerscijournal.org/vol8no1/a-comparative-study-of-classification-techniques-in-data-mining-algorithms/

[16].A Comparative Study of Classification Techniques in Data Mining Algorithms

Sagar S. Nikam, Department of Computer Science, K.K.Wagh College of Agriculture, Nashik, India.

[17].https://www.kdnuggets.com/2015/05/top-10-data-mining-algorithms-explained.html

[18].http://playwidtech.blogspot.com/2013/02/k-means-clustering-advantages-and.html

[19].https://dhirajkumarblog.medium.com/top-4-advantages-and-disadvantages-of-support-vector-machine-or-svm-a3c06a2b107