JCSE International Journal of Computer Sciences and Engineering Open Access

Review Paper Volume-4, Issue-12 E-ISSN: 2347-2693

# A Review: Comparative Analysis of various Data Mining Techniques

**Pinki Sagar, Monika Goyal***

[1,2]Dept. of Computer Science and Engineering, Manav Rachna International University, India

**Available online at: www.ijcseonline.org**

Received: 26/Nov/2016 Revised: 03/Dec/2016 Accepted: 20/Dec/2016 Published: 31/Dec/2016

**Abstract*:*** D*ata mining* (sometimes called *data* or knowledge discovery) is the process of analyzing *data* from different perspectives and summarizing it into useful information – making it more accurate, reliable, efficient and beneficial. In data mining various techniques are used- classification, clustering, regression, association mining. These techniques can be used on various types of data; it may be stream data, one dimensional, two dimensional or multi-dimensional data. In this paper we analyze the data mining techniques based on various parameters. All data mining techniques used for prediction, extraction of useful data from a large data base. Each of the techniques have different performance and result .

***Keywords: Data mining, Classifications,Prediction,Clustering,Associatio***

## I. Introduction:

The Data Mining Specialization teaches data mining techniques for both structured data which conform to a clearly defined schema, and unstructured data which exist in the form of natural language text. Specific course topics include pattern discovery, clustering, text retrieval, text mining and analytics, and data visualization. Database Analysis of data in a database using tools which look for trends or anomalies without knowledge of the meaning of the data. Data Mining is defined as the procedure of extracting information from huge sets of data. In other words, we can say that data mining is mining knowledge from data.

We have broken the discussion into two sections, each with a specific theme:

- Classical Techniques: Statistics, Neighborhoods and Clustering
- Next Generation Techniques: Trees, Networks and Rules

Each section will describe a number of data mining algorithms at a high level, focusing on the "big picture" so that the reader will be able to understand how each algorithm fits into the landscape of data mining techniques.
Overall, six broad classes of data mining algorithms are covered. Although there are a number of other algorithms and many variations of the techniques described, one of the algorithms from this group of six is almost always used in real world deployments of data mining systems.

Our paper is organized as follows:
Section II classifies the various data mining techniques in detail.
Section III lists various application areas of Data mining.
Section IV differentiates Prediction from Classification.
Section V shows analysis of various data mining techniques in a tabular form.

## II. Data Mining Algorithms and Techniques

Various algorithms and techniques like Classification, Clustering, Regression, Artificial Intelligence, Neural Networks, Association Rules, Decision Trees, Genetic Algorithm, Nearest Neighbor method etc., are used for knowledge discovery from databases.

### Classification:

Classification is the most commonly applied data mining technique, which employs a set of pre-classified examples to develop a model that can classify the population of records at large. Fraud detection and credit- risk applications are particularly well suited to this type of analysis. This approach frequently employs decision tree or neural network-based classification algorithms. The data classification process involves learning and classification. In Learning the training data are analyzed by classification algorithm. In classification test data are used to estimate the accuracy of the classification rules. If the accuracy is acceptable the rules can be applied to the new data tuples. For a fraud detection application, this would include complete records of both fraudulent and valid activities determined on a record-by-record basis. The classifier-training algorithm uses these pre-classified examples to determine the set of parameters required for proper discrimination. The algorithm then encodes these parameters into a model called a classifier.

Types of classification models:

- Classification by decision tree induction
- Bayesian Classification
- Neural Networks
- Support Vector Machines (SVM)
- Classification Based on Association

**Clustering:**

Clustering can be said as identification of similar classes of objects. By using clustering techniques we can further identify dense and sparse regions in object space and can discover overall distribution pattern and correlations among data attributes. Classification approach can also be used for effective means of distinguishing groups or classes of object but it becomes costly so clustering can be used as preprocessing approach for attribute subset selection and classification. For example, to form group of customers based on purchasing patterns, to categories genes with similar functionality.

Types of clustering methods

- Partitioning Methods
- Hierarchical Agglomerative (divisive) methods
- Density based methods
- Grid-based methods
- Model-based methods

**Prediction:**

Regression technique can be adapted for predication. Regression analysis can be used to model the relationship between one or more independent variables and dependent variables. In data mining independent variables are attributes already known and response variables are what we want to predict. Unfortunately, many real-world problems are not simply prediction. For instance, sales volumes, stock prices, and product failure rates are all very difficult to predict because they may depend on complex interactions of multiple predictor variables. Therefore, more complex techniques (e.g., logistic regression, decision trees, or neural nets) may be necessary to forecast future values. The same model types can often be used for both regression and classification. For example, the CART (Classification and Regression Trees) decision tree algorithm can be used to build both classification trees (to classify categorical response variables) and regression trees (to forecast continuous response variables). Neural networks too can create both classification and regression models.

Types of regression methods

- Linear Regression
- Multivariate Linear Regression
- Nonlinear Regression
- Multivariate Nonlinear Regression

**Association rule**

Association and correlation is usually to find frequent item set findings among large data sets. This type of finding helps businesses to make certain decisions, such as catalogue design, cross marketing and customer shopping behavior analysis. Association Rule algorithms need to be able to generate rules with confidence values less than one. However the number of possible Association Rules for a given dataset is generally very large and a high proportion of the rules are usually of little (if any) value. Types of association rule

- Multilevel association rule
- Multidimensional association rule
- Quantitative association rule

**Neural networks :**

Neural network is a set of connected input/output units and each connection has a weight present with it. During the learning phase, network learns by adjusting weights so as to be able to predict the correct class labels of the input tuples. Neural networks have the Remarkable ability to derive meaning from complicated or imprecise data and can be used to extract patterns and detect trends that are too complex to be noticed by either humans or other computer techniques. These are well suited for continuous valued inputs and outputs. For example handwritten character reorganization, for training a computer to pronounce English text and many real world business problems and have already been successfully applied in many industries. Neural networks are best at identifying patterns or trends in data and well suited for prediction or forecasting needs.

**Prediction:**

There are two forms of data analysis that can be used for extracting models describing important classes or to predict future data trends. These two forms are as follows −

- Classification
- Prediction

Classification models predict categorical class labels; and prediction models predict continuous valued functions. For example, we can build a classification model to categorize bank loan applications as either safe or risky, or a prediction model to predict the expenditures in dollars of potential customers on computer equipment given their income and occupation. Predicting the identity of one thing based purely on the description of another, related thing

• Not necessarily future events, just unknowns
• Based on the relationship between a thing that you can know and a thing you need to predict.

## III. Data Mining Applications

Here is the list of areas where data mining is widely used −

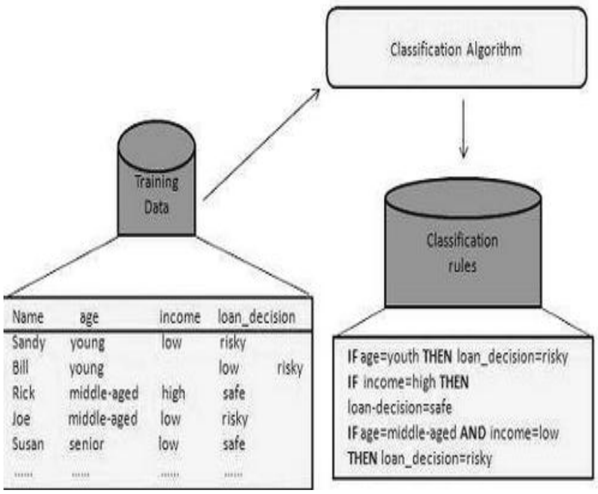
- Financial Data Analysis
- Retail Industry
- Telecommunication Industry
- Biological Data Analysis
- Other Scientific Applications
- Intrusion Detection

## IV How Prediction Differs From Classification

• A classification problem could be seen as a predictor of classes, but .
• Predicted values are usually continuous whereas classifications are discreet.
• Predictions are often (but not always) about the future whereas classifications are about the present.
• Classification is more concerned with the input than the output

**Classification**: Classification process includes following steps −

- Building the Classifier or Model
- Using Classifier for Classification
- This step is the learning step or the learning phase.
- In this step the classification algorithms build the classifier.
- The classifier is built from the training set made up of database tuples and their associated class labels.
- Each tuple that constitutes the training set is referred to as a category or class. These tuples can also be referred to as sample, object or data points

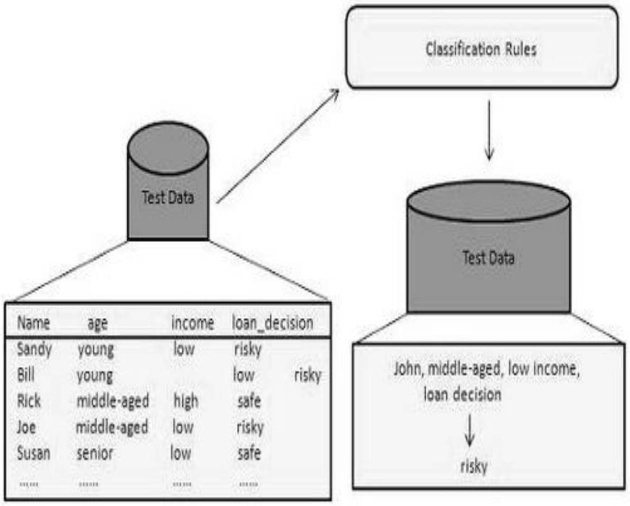


| Name | age | income | loan_decision | |
|---|---|---|---|---|
| Sandy | young | low | risky | |
| Bill | young | | low | risky |
| Rick | middle-aged | high | safe | |
| Joe | middle-aged | low | risky | |
| Susan | senior | low | safe | |
| ...... | ...... | ...... | ...... | |

IF age=youth THEN loan_decision=risky
IF income=high THEN
loan-decision=safe
IF age=middle-aged AND income=low
THEN loan_decision=risky

**Figure 1:Using Classifier for Classification**

In this step, the classifier is used for classification. Here the test data is used to estimate the accuracy of classification rules. The classification rules can be applied to the new data tuples if the accuracy is considered acceptable.



| Name | age | income | loan_decision | |
|---|---|---|---|---|
| Sandy | young | low | risky | |
| Bill | young | | low | risky |
| Rick | middle-aged | high | safe | |
| Joe | middle-aged | low | risky | |
| Susan | senior | low | safe | |
| ...... | ...... | ...... | ...... | |

John, middle-aged, low income,
loan decision
↓
risky

Figure2: classification

**Association :**

Basic terminology:

1. Tuples are *transactions*, attribute-value pairs are *items.*
2. *Association rule*: {A,B,C,D,...} => {E,F,G,...}, where A,B,C,D,E,F,G,... are items.

3. *Confidence* (accuracy) of A => B : P(B|A) = (# of transactions containing both A and B) / (# of transactions containing A).
4. *Support* (coverage) of A => B : P(A,B) = (# of transactions containing both A and B) / (total # of transactions)
5. We looking for rules that exceed pre-defined support (*minimum support*) and have high confidence.

Example:

| **Transaction id** | **Items** |
|---|---|
| T1 | Bread, Jelly, PeanutButter |
| T2 | Bread, PeanutButter |
| T3 | Bread, Milk , PeanutButter |
| T4 | Beer, Bread |
| T5 | Beer, Milk |

Table 1:Database for Transactions

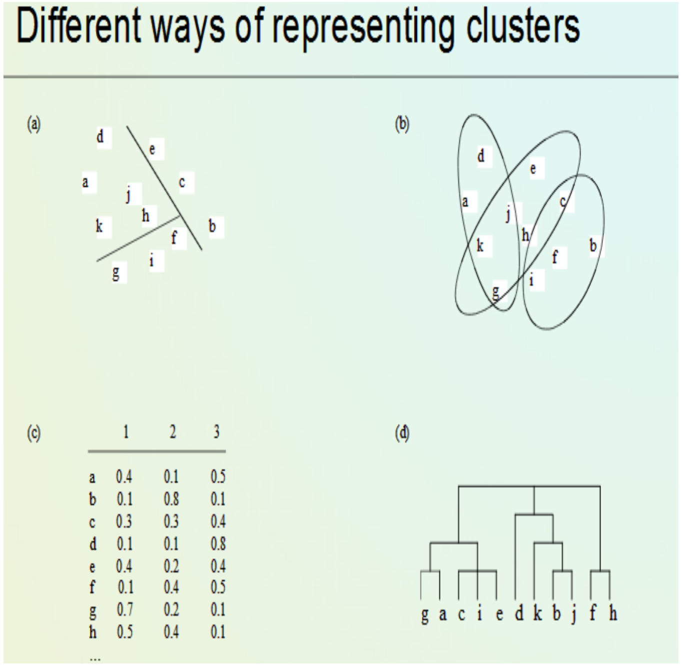
| **X => Y** | **s(support )** | **a(confidence )** |
|---|---|---|
| Bread=> PeanutButter | 60% | 75% |
| PeanutButter=> Bread | 60% | 100% |
| Beer=>Bread | 20% | 50% |
| PeanutButter=>Jelly | 20% | 33.3% |
| Jelly=>PeanutButter | 20% | 100% |
| Jelly=>Milk | 0% | 0% |

Table 2: Association rules for transactions

**Clustering:**A cluster is a subset of objects which are “similar”

1. A subset of objects such that the distance between any two objects in the cluster is less than the distance between any objectin the cluster and any object not located inside it.
2. A connected region of a multidimensional space containing a relatively high density of objects.

Clustering is a process of partitioning a set of data (or objects) into a set of meaningful sub-classes, called clusters
.
• Help users understand the natural grouping or structure in a Clustering: unsupervised classification no predefined classes.
• Used either as a stand-alone tool
to get insight into data distribution or as a preprocessing step for other algorithms.
• Moreover, data compression, outlier’s detection, understands human concept formation.



Different ways of representing clusters

|   | 1 | 2 | 3 |
|---|---|---|---|
| a | 0.4 | 0.1 | 0.5 |
| b | 0.1 | 0.8 | 0.1 |
| c | 0.3 | 0.3 | 0.4 |
| d | 0.1 | 0.1 | 0.8 |
| e | 0.4 | 0.2 | 0.4 |
| f | 0.1 | 0.4 | 0.5 |
| g | 0.7 | 0.2 | 0.1 |
| h | 0.5 | 0.4 | 0.1 |
| ... |   |   |   |

Figure 3: Clustering method.

### V Analysis Table

| *Parameters* | *Classification* | *Association* | *Regression* | *Clustering* |
|---|---|---|---|---|
| *Performance* | High | Moderate | High | Moderate |
| *Technique Complexity* | High | Low | Very High | Low |
| *Requirement of Dependent Variable* | no | yes | yes | No |
| *Algorithms* | ID3 algorithm,C4.5 algorithm,SLIQ algorithm |  | FIPM<br>FTPDS<br>Linear<br>Nonlinear regression | K-MEANS<br>K-MODES |

| | | | | |
|---|---|---|---|---|
| *Type of data* | Two dimensional data | One dimensional data | One dimensional and two dimensional data | One dimensional and Multi -dimensional data |
| *Applications* | Expert Systems and statistics Neuron Biology | Market Basket Analysis, frequent pattern searching, Cross Marketing | Deviation Detection | Statistics biology machine learning |
| *Analysis* | Discriminate | Specific | Calculation of errors during the prediction | Exploratory |
| *Supervised /Unsupervised learning* | Supervised | Supervised and unsupervised | Supervised | Unsupervised |

Table 4: Analysis Table for data mining techniques

## References

[1] L. Breiman, J.H. Friedman, R.A. Olshen, and C.T. Stone. Classification and Regression Trees. Wadsworth, Belmont, California, 1984.

[2] C.L. Blake D.J. Newman, S. Hettich and C.J. Merz. UCI reposito ry of machine learning databases, 1998.

[3] T. Elomaa and M. K¨a¨ari¨ainen. An analysis of reduced error pruning.Journal of Artificial ntelligence Research , 15:163–187, 2001.

[4] Usama M. Fayyad. Data mining and knowledge discovery: Makin g sense out of data. IEEE Expert: Intelligent Systems and Their Applications 11(5):20–25, 1996.

[5] A. Feelders. classification trees. ttp://www.cs.uu.nl/docs/ vakken/adm/trees.pdf.

[6] R. Kruse G. Della Riccia and H. Lenz.Computational Intelligence in Data Mining. Springer, New York, NY, USA, 2000.

[8] J. Ross Quinlan. C4.5: programs for machine learning. Morgan Kauf-mann Publishers Inc., San Francisco, CA, USA, 1993.

[9] Ian H. Witten and Eibe Frank.Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann Publishers Inc., San francisco,CA, USA, 2nd edition, 2005.

[10] D.F. Andrews, :A robust method for multiple linear ression,T*echnometrics* , vol 16, 1974, pp 125 – 127

[11]Chai, Eun Hee Kim and Long Jin:prediction of Frequent Items to OneDimensional Stream Data; Fifth International Conference on Computational Science and Applications ; page 353-360, 2001

[12]Y. Chen, G.Dong, J.Han, B.W.Wah, and J.Wang :.Multi-Dimensional Regression Analysis of Time- Series Data Streams; Proc. Int. Conf. Very Large Data Bases;Hong Kong, China, Aug. 2002.

[13] “Regression Based Data Mining Techniques for Frequent Data Stream *(One Dimensional and two Dimensional Stream Data”* Pinki Sagar, International Journal of Computer Sciences and Engineering pp(140-143), Vlume3, Issue 9.,2015

[14] “Improving the Initial Centroids of K-Means Clustering Algorithm to Generalize its Applicability”,M. Goyal, s. Kumar,Journal of TheInstitution of Engineers December 2014, Volume 95, Issue 4, pp 345–350,

## Authors Profile

*Ms. Pinki Sagar* pursed Bachelor of Technology from YMCA university , india in 2006 and Master of Technology from **Mahirishi Dayanand University** in year 2009. SHe is currently pursuing Ph.D. and currently working as Assistant Professor in Department of Computer Science and engineering, Manav Rachna International University since 2006. She has published more than 10 research papers in data mining and prediction. in reputed international journals and conferences.. SHe has 10 years of teaching Experience .

*Ms.Monika Goyal* pursed Bachelor of Technology from kurukshetra university , india in 2005 and Master of Technology from **Mahirishi Dayanand University** in year 2011. Currently She is working as Assistant Professor in Department of Computer Science and engineering.Manav Rachna International University since 2008. Her research work has been published in reputed jounal **SPRINGER**. She has 10 years of teaching Experience .