Analysis of Prediction Techniques based on Classification and Regression

Pinki Sagar
Research Scholar
MRU,Faridabad

Prinima, PhD
Assistant Professor
MRU, Faridabad

Indu, PhD
Associate Professor
MRIU,Faridabad

ABSTRACT

Data 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 in various fields for prediction and extraction of useful data or knowledge from a large data base is analyzed and each data mining technique has different performance.

General Terms

Data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information. c*lassification* is a general process related to categorization, the process in which ideas and objects are recognized, differentiated, and understood. A *classification* system is an approach to accomplishing *classification*.

Keywords

Data mining, Classification, Prediction, Clustering, Association

1. 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. Analysis of data in a database using tools which looks 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 of knowledge from data. The tutorial starts off with a basic overview and the terminologies involved in data mining and then gradually. Moves on to cover topics such as knowledge discovery, query language, classification and Prediction, decision tree induction, cluster analysis, and how to mine the Web. 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 data mining technique 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.

2. 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.

3. 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 [3]. 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 [9]. The algorithm then encodes these parameters into a model called a classifier.

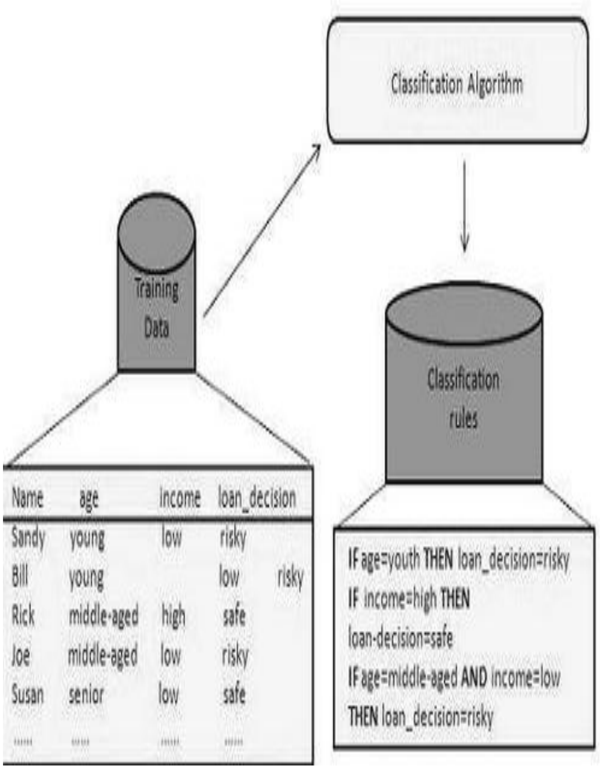
Various types of classification models are:

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

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

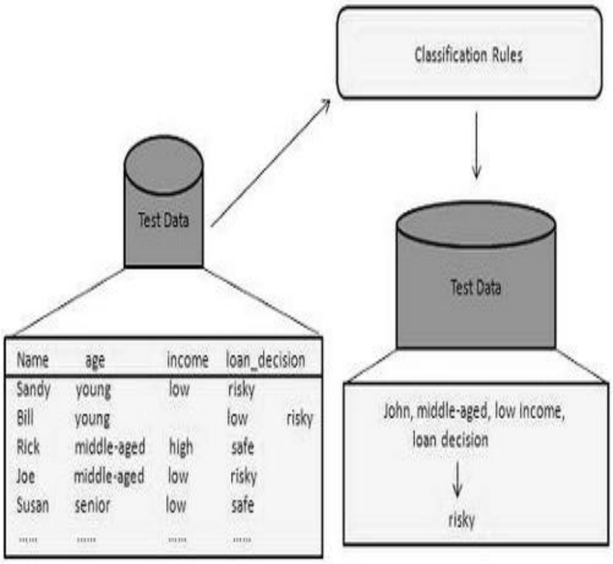


Name	age	income	loan_decision	
Sandy	young	low	risky	
Bill	young		low	risky
Rick	middle-aged	high	safe	
Joe	middle-aged	low	risky	
Susan	senior	low	safe	
.....	.....	.....	.....	

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

Figure 2: Classification

4. 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[4]. For example, to form group of customers based on purchasing patterns, to categories genes with similar functionality.

Various types of clustering methods are:

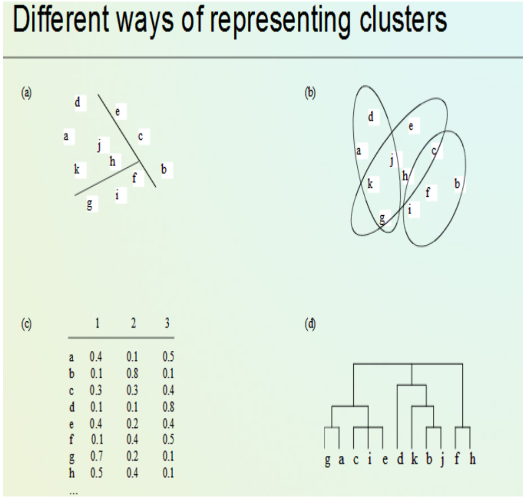
- Partitioning Methods
- Hierarchical Agglomerative (divisive) methods
- Density based methods
- Grid-based methods
- Model-based methods

4.1 Clustering

A cluster is a subset of objects which are “similar”. A subset of objects such that the distance between any two objects in the cluster is less than the distance between any object in the cluster and any object not located inside it. A connected region of a multi-dimensional 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. Clustering helps in following ways:

- 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

5. REGRESSION

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 [11]. 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.

Various types of Regression methods are:

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

5.1 Step for FIPM (Frequent Item Prediction Method) are:

Step1: Reorganize the stream data according to the occurrence of particular time.

Step2: Calculate the time difference between the previous time and next time of variable when the data arrived.

Step3: create the pairs from the array of difference of time

for example:

- **(x1,x2,x3,x4)**: {x1andx2},{x2and x3),{x3andx4}
- **(a1,a2,a3,a4)** : (a1,a2),(a2,a3),(a3,a4)

Variable of x independent variable and y dependent variable
X: a1,a2,a3

Y: a2,a3,a4

Step4: calculation of co efficient of regression model of

FIPM b0 and b1

$$b_0 = \bar{y} - b_1 \bar{x}$$

$$\hat{b_i} = \frac{\sum x_i y_i - n\overline{xy}}{\sum x_i^2 - nx^{-2}} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

Step5: fit the regression model of FIPM:

$$\dot{y} = b_0 + b_1 x + E$$

5.2 SFAPR: Preprocessing of training data for Sequence Forecast Algorithm on Plane Regression algorithm [11]

Calculate the support sequence for particular data according to their ids, time and sliding windows for any specified stream data sequence. Sliding windows are 0-3, 1-4, 2-5, 3-6 and so on. Support or actual F(frequency for appearing the stream data) is calculated using the following method. independent variables, and then For SFA-PR We calculate the coefficient using preprocessing and fit the regression model. For calculating the coefficients and regression model we used the various symbols:

Σf	Sum of all frequencies (or support) Dependent variables (1,2----n)
Σt	Sum of all times independent Variables (1, 2 -------n)
Σtf	Sum of multiplication of time and Frequencies (1, 2, -------n)

For calculating the regression model we use the following equations so that we can predict the frequency at which data is appearing.

SFA-PR the regression model is:

Y= b0+**b1t**+b2f+□ □ (5)

There exists constant n and matrix Y, X, and β, let

F= Num of id’s at which stream data is presented/

Total number of id’s of stream data

$$Y = \begin{bmatrix} Y1 \\ Y2 \\ . \\ . \\ . \\ YN \end{bmatrix}$$

$$B = \begin{bmatrix} B0 \\ B1 \\ B2 \end{bmatrix}$$

$$X = \begin{bmatrix} 1 & t1 & f1 \\ 1 & t2 & f2 \\ 1 & tn & fn \end{bmatrix}$$

$$X'x' = \begin{bmatrix} n & \sum t & \sum f \\ \sum t & \sum t^2 & \sum tf \\ \sum f & \sum tf & \sum f^2 \end{bmatrix}$$

$$X'y = \begin{bmatrix} \sum y \\ \sum y \\ \sum fy \end{bmatrix}$$

6. 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

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.

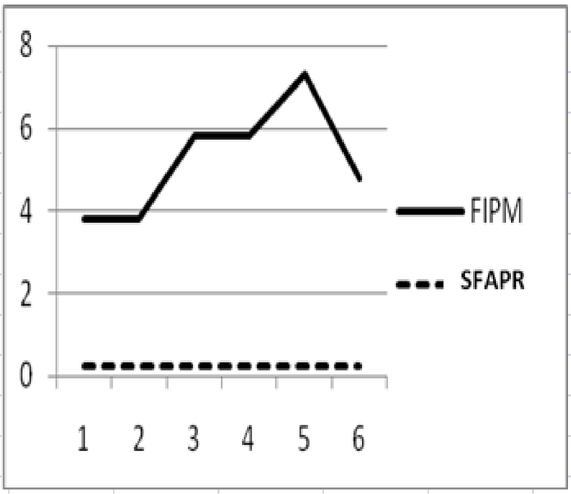
X => Y	**s(support)**	**a(confidence)**
Bread=>PeanutButter	60%	75%
PeanutButter=> Bread	60%	100%
Beer=>Bread	20%	50%
PeanutButter=>Jelly	20%	33.3%
Jelly=>PeanutButter	20%	100%
Jelly=>Milk	0%	0%

Association rules for transactions

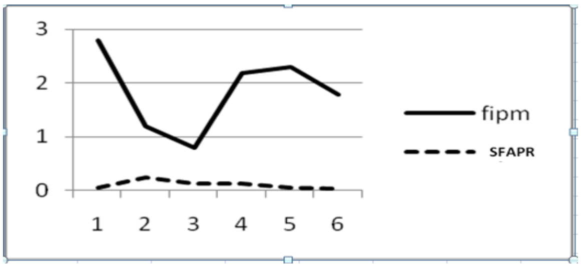
6.1 Analysis Table

Parameters	**Classification**	**Regression**
Performance	**High**	**High**
Technique Complexity	**High**	**Very High**
Requirement of Dependent Variable	**No**	**Yes**
Algorithms	**ID3 algorithm,C4.5 algorithm,SLIQ algorithm**	**FIPM** **FTPDS** **Linear** **Nonlinear regression**
Type of data	**Two dimensional data**	**One dimensional and two dimensional data**
Applications	**Expert Systems and statistics Neuron Biology**	**Deviation Detection**
Analysis	**Discriminate**	**Calculation of errors during the prediction**
Supervised /Unsupervised learning	**Supervised**	**Supervised**

Parameters	**FIPM**	**FTP-DS**	**SFAPR**
Type of Data	**One dimensional Stream Data**	**Two Dimensional Stream Data**	**For Multidimensional Data Stream**
Error	**High error**	**Low error**	**Very low**
Method	**Linear and Non Linear Method**	**Linear and Non Linear Method**	**Non Linear Method**
Accuracy	**low**	**High**	



Prediction curve for FIPM and SFA-PR



Error curve for FIPM and SFA-PR

7. CONCLUSION

In this paper we analyze that in data mining, prediction of data can be done using two data mining techniques classification and regression. These data mining techniques have various algorithms. In prediction using classification accuracy is low with high errors, but using regression we can get accurate prediction of data with low errors. We compared classification and regression algorithms under the various parameters like accuracy of prediction, complexity, errors etc. After comparing these two techniques we can say that ,efficient prediction can be done using regression and its algorithms.

8. 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 repository of machine learning databases, 1998.

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

[4] Usama M. Fayyad. Data mining and knowledge discovery: Makin g senseout 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 inData Mining. Springer, New York, NY, USA, 2000.

[7] N. Landwehr, M. Hall, and E. Frank. Logistic model trees, 2003

[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 EibeFrank.Data Mining: Practical machine learningtools and techniques. Morgan Kaufmann Publishers Inc., San francisco,CA, USA, 2nd edition, 2005.

[10] D.F. Andrews, :A robust method for multiple linear regression,T*echnometrics*, vol16,1974,pp125–127

[11] Chai, EunHeeKim and Long Jin:predictionof Frequent Items to OneDimensionalStream Data; Fifth International Conference on Computational Science and Applications; page353-360,2001

[12] Y. Chen, G.Dong, J.Han, B.W.Wah, andJ.Wang:.Multi dimensionalRegressionAnalysisofTime-Series DataStreams; Proc.Int.Conf.Very LargeDataBases;HongKong,China, Aug.2002.

[13] R. Hayward; A Basic Approach to Linear Regression; RWJ linical Scholars Program; pp1-3,University of Michigan , 2005.

[14] O.B.Yaik, C.H.Yong, and FHaron, Time Series Prediction using Adaptive Association rules,InProc.of DFMA05, pp.310-314, 2005.

[15] Omid Rouhani-Kalleh; Algorithms for Fast Large Scale data Mining Using Logistic Regression; Proceedings of the 2007 IEEE Symposium on Computational Intelligence and Data Mining; pp 155-162,2007

[16] Feng Zhao, Qing-Hua A Li :A Plane Regression Based Sequence Forecast Algorithms for Stream Data ; Proc. of the Fourth International Conference on Machine Learning and Cybernetics; pp-1559-1562 Guangzhou,18-21August, 2005.

[17] Y. Peng, G. Kou, Y. Shi, Z. Chen; A Descriptive Framework for the Field of Data Mining and Knowledge Discovery. International Journal of Information Technology and Decision Making, Volume 7, Issue 4: 639 – 682; 2000

[18] Perlich, C,Provost, F., Simonoff, J. S. TreeInduction verses. Logistic Regression:A Learning-Curve

Analysis. Journal of Machine Learning Research Vol. 4 pp-211- 255. 2003.

[19] Amir Bar-Or, Daniel Keren, Assaf Schuster, and Ran Wolff: Hierarchical Decision Tree Induction in istributed Genomic Databases; IEEERANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING,VOL. 17;pp;1138- 1150,2007.

[20] Qi Luo; Advancing Knowledge Discovery and Data Mining; Workshop on Knowledge Discovery and Data Mining pp;3-5, 2008.

[21] Fayyad, Usama; Gregory Piatetsky-Shapiro, and adhraic Smyth; From Data Mining to Knowledge Discovery in Databases. -pp:12-17, June 2008.