ISSN 2229-3531 (Print)
JUCIT Vol. 8(6), 37-40 (2017). Periodicity-2-Monthly
ISSN 2455-9997 (Online)

JOURNAL OF COMPUTER & INFORMATION TECHNOLOGY
An International Open Free Access Peer Reviewed Research Journal of Computer Science Engineering & Information Technology
website:- www.compitjournal.org
Estd. 2010

A comparative Analysis of Multiple Regression in Data Mining

PRIYANKA VERMA [1], RAJNI KORI[2] and SHIV KUMAR[3]

M. Tech Scholar, Department of Computer Science & Engineering, Lakshmi Narain College of Technology & Excellence Bhopal (M.P), (India), Email: roshni180393@gmail.com[1]

Assistant Professor, Department of CSE, Lakshmi Narain College of Technology & Excellence, Bhopal (M.P), (India)[2]

Professor & Head, Department of CSE, Lakshmi Narain College of Technology & Excellence, Bhopal (M.P), (India)[3]

http://dx.doi.org/10.22147/jucit/080601

Acceptance Date 10th November, 2017, Online Publication Date 02nd December, 2017

Abstract

The growing volume of data usually creates an interesting challenge for the need of data analysis tools that discover regularities in these data. Data mining has emerged as disciplines that contribute tools for data analysis, discovery of hidden knowledge, and autonomous decision making in many application domains. The Multiple regressions generally explain the relationship between multiple independent or multiple predictor variables and one dependent or criterion variable. The regression algorithm estimates the value of the target (response) as a function of the predictors for each case in the build data. These relationships between predictors and target are summarized in a model, which can then be applied to a different data set in which the target values are unknown.

In this paper, we have discussed the formulation of multiple regression technique, along with that multiple regression algorithm have been designed, further test data are taken to prove the multiple regression algorithm.

Key words: Multiple regression, dependent variable, independent variables, predictor variable, response variable

I Introduction

Social Predictive modelling is a name given to a collection of mathematical techniques having in common the goal of finding a mathematical relationship between a target, response, or "dependent" variable and various predictor or "independent" variables with the goal in mind of measuring future values of those predictors and inserting them into the mathematical relationship to predict future values of the target variable, it is desirable to give some measure of uncertainty for the predictions, typically a prediction interval that has some assigned level of confidence like 95%.Regression analysis establishes a relationship between a dependent o r outcome variable and a set of predictors. Regression, as a data mining technique, is supervised learning. Supervised learning par t i t ions the database into training and validation data. The techniques used in this research were simple linear regression and multiple linear regression. Some distinctions between the uses of regression in statistics verses data mining are: in statistics the data is a sample from a population, but in Data Mining the data is taken from a large database (e.g. 1 million records). Also in statistics the regression model is constructed from a sample, but in Data Mining the regression model is constructed from a portion of the data (training data). Predictive analytics encompasses a variety of

techniques from statistics, data mining and game theory that analyze current and historical facts to make predictions about future events. The variety of techniques is usually divided in three categories: predictive models, descriptive models and decision models.

Predictive models look for certain relationships and patterns that usually lead to a certain behaviour, point to fraud, predict system failures, assess credit worthiness, and so forth. By determining the explanatory variables, you can predict outcomes in the dependent variables.

Descriptive models aim at creating segmentations, most often used to classify customers based on for instance socio-demographic characteristics, life cycle, profitability, product preferences and so forth. Where predictive models focus on a specific event or behaviour, descriptive models identify as many different relationships as possible.

Decision models that use optimization techniques to predict results of decisions. This branch of predictive analytics leans particularly heavily on operations research, including areas such as resource optimization, route planning and so forth.

1.1 Data Mining Techniques :

Knowledge or Information for decision making in a business is very poor even though data storage grows exponentially. Data mining also known as Knowledge discovery. The Knowledge extracted allows predicting the behaviour and future behaviour. This allows the business owners to take positive, knowledge driven decisions. Data mining is applied on various industries like retail, finance, health care, aerospace, education etc. Knowledge is extracted from the historical data by applying pattern recognition, statistical and mathematical techniques those results in the knowledge the form of facts, trends, association, patterns, anomalies and exceptions. There are some areas where data mining will be applied.

Data Pre-processing: Data pre-processing make ready the real world data for mining process.

Data Mining: data mining is the process of extracting some important patterns from a large amount of data.

Pattern Evaluation: This process evaluates the pattern that is generated by the data mining. The patterns are evaluated according to the interestingness measure given by user or system.

Knowledge presentation: Knowledge Presentation uses visualization techniques that visualize the interesting patterns and help the user to understand and interpret the resultant patterns.

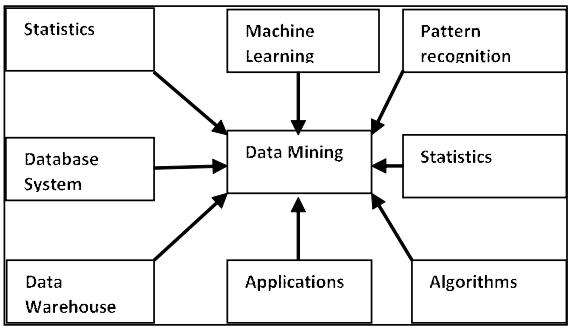

Figure 1: Application of Data mining

1.2 Overview of data mining :

Data mining has attracted a great deal of attention in the information industry and in society as a whole in recent years, due to availability of large amount of data and imminent need for turning such data into useful information and knowledge. Data mining is the process of digging through data and looking meaningful trends and patterns. The information and knowledge gained can be used for applications ranging from market analysis, fraud detection, and customer retention, to production control and science exploration Data mining can be viewed as a result of the natural evolution of information technology. Data mining is iterative process.

Data cleaning: It is a process of removing noise and inconsistent data.

Data integration: In this step data from multiple sources are combined.

Data selection: In this step data relevant for mining task is selected.

Data transformation: In this step data will be transformed into form that is appropriate for mining.

Data mining: In this step some intelligent methods are applied for extracting data patterns.

Pattern evaluation: In this step truly interesting patterns representing knowledge based on some interestingness measure are identified.

Knowledge presentation: In this step visualization and knowledge representation techniques are used to present the mined knowledge to the user.

1.3 Data Mining Algorithms & Techniques :

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

1.4 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.

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

1.5 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 by decision tree induction

- Partitioning Method
- Hierarchical Agglomerative methods
- Density based methods
- Grid-based methods
- Model-based methods

1.6 Prediction :

Clustering 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

II Literature Survey :

2.1 Forecasting is analysing and predicting the future behaviour:

In this paper Author conclude that Forecasting is analysing and predicting the future behaviour of the selected data set. In Forecasting knowledge from the analysed data is used to predict future behaviours. It help in many ways in various domains such as controlling load balance, future marketing campaigns, allocating or de-allocating resources and caching , perfecting web pages for improve performance. There are limited number of researches have been done on Web site related forecasting.

2.2 Social Media Mining For Public Health Monitoring and Surveillance :

This paper describes topics Neha Khandelwal, *et al.*[1] presented a MLR equation to predict rainfall using four different climatic factors for Jaipur city, Rajasthan, India, for selecting these factors the author used Pearson correlation coefficient and then use the result to determine the drought possibility.

III Problem Identification :

A lot of research work has been done in the field. Authors have learned several things from this study (work). First, we will introduce some terms that are related with our topic such as, some brief description of Big Data, Data Warehouses, Data Mining and their classification and finally also will do the particular analysis for Regression technique (linear and multiple regression) and our approach regarding a prototype and how can be used and for what reasons regression techniques giving explanations with concrete examples. Regression as technique although is predictive technique, but based on analyzes conducted to reach the conclusion most scientists, they have concluded that the reliability percentage is around 95%. Through our analysis we will try to demonstrate this scale of reliability through concrete examples.

IV Block Diagram & Methodology :

4.2 Algorithm :

1. Input / Load data set
2. Apply feature extraction
3. Received Extracted data as output
4. Generate Training and Testing data set (By applying

techniques:)

5. Apply Multiple Linear Regression algorithm to training dataset (MLR)
6. Build the Reduction Explanatory Predictor
7. Building Model using Regression
8. Perform / Obtain validity check
9. Utilize the "test" set predictions to calculate all the performance metrics (Measure Accuracy and other parameters)

V Conclusion

The Multiple Regression technique predicts a numerical value. Regression performs operations on a dataset where the target values have been defined already, and the result can be extended by adding new information. The relations which regression establishes between predictor and target values can make a pattern. This pattern can be used on other datasets where the target values are not known. In this paper we have formulate a multiple regression technique, further we have designed the multiple regression algorithm. The test data are taken to prove the relationship between predictor and target variable which is being represented by the linear regression equation

$$Y= \alpha + \beta 1\ X1+ \beta 2\ X2$$

where random variable Y (called as a response variable) as a linear function of random variable X1 (called as a predictor variable) and X2 α and β are linear regression coefficients.

References

1. Manisha rathi Regression modeling technique on data mining for prediction of CRM CCIS 101, pp.195-200, 2010 Springer–Verlag Heidelberg (2010).
2. Jiawei Han and Micheline Kamber, Data Mining Concepts and Techniques, published by Morgan Kauffman, 2nd ed. (2006).
3. Giudici Paolo, "Applied Data Mining-Statistical methods for business and industry" wiley, (2003) [5] Dash, M., and H. Liu, "Feature Selection for Classification," Intelligent Data Analysis. 1:3 (1997) pp. 131-156. [6] Rencher C. Alvin, "Methods of Multivariate Analysis" 2nd Edition, Wiley Interscience (2002).
4. Burges, C., A tutorial on Support Vector Machines for Pattern Recognition. Data Mining and Knowledge Discovery, *2(2)*, 955–974 (1998).
5. CiteSeer, CiteSeer Scientific Digital Library. http://www.citeseer.com. (2002).
6. Duda, R. O. and Hart, P. E., Pattern Classification and Scene Analysis. John Wiley & Sons. (1973).
7. Greenbaum, A., Iterative Methods for Solving Linear Systems, volume 17 of Frontiers in Applied Mathematics. SIAM (1997).
8. GLIM, Generalised Linear Interactive Modelingpackage.http://www.nag.co.uk/stats/GDGEsoft.asp, http://lib.stat.cmu.edu/glim/. (2004).
9. Kubica, J., Goldenberg, A., Komarek, P., Moore, A., and Schneider, J., A comparison of statistical and machine learning algorithms on the task of link completion. In KDD Workshop on Link Analysis for Detecting Complex Behavior, page 8 (2003).
10. Lay, D. C., Linear Algebra and Its Applications. Addison-Wesley (1994).