Regression Analysis in Data Mining – A Survey

Anuj Pauranik#1, Deepak Pathak*2, Gagan Sharma#3 Dr. A.C.Nayak#4

*#1MTech Scholar, *2Assistant Professor, #3Assistant Professor, #4Assistant Professor*

RKDF University, Gandhi nagar Bhopal, India

1anuj87in@gmail.com

2 deep_325@yahoo.com

3 gagansharma.cs@gmail.com

4 dracnayak@ymail.com

Abstract — Various sources ranging from medical data to social media and corporate data has resulted in rapid growth of data. Such a huge collection of data is known to be "Big Data". The difficult and challenging part is to analyse and process such a large amount of data from various sources is to extract useful and fruitful information from such a big pool of data. Conventional data analysis technique are not much capable in respect to big data to the varying size of the data, limited memory and slow processors.. In this paper, we have presented a survey on effective Regression Analysis for big data. We shall be discussing different aspects of Regression analysis with their pros and cons.

Keywords — Regression, Big Data, Map Reduce, Data Mining

1. Introduction

The term Big data is referred to the data that is very large, complex and varying in nature thus making it difficult to process for finding any conclusion or fruitful outcome out of it. Conventional processing methods fall short to do that. The area has gained momentum in early 2000. The Internet era presents a big volume of data where large amounts of information are generating every second. "Apache Hadoop" is a collection of open source software utilities to resolve the issues in large sets of data having difference in variety, velocity and veracity. Hadoop can reduce huge process duration to a few hours or minutes using MapReduce as a parallel processing mechanism.

Data Mining is a very vast area to be studied and analyzed. A very vital role is played by statistics. It is a form of mathematical Analysis that uses representations and models for real life data studies. So there exist many statistical methods for data analysis. They offer wide range of features for data modeling and analysis. Regression is found in so many fields like schools, university, social media, medical data, business, scientific researches, weather forecast, human behavior and so on. Regression Analysis is an important aspect for data analysis in the above mentioned fields for various purposes like student records, prediction of rain or storms, people's purchasing behavior for business analysts, social media behavior and by various corporate houses and business owners for analysis and prediction of profit and loss in their business.

2. Regression Analysis

Regression analysis [7] is a tool that is statistical in nature which helps in establishing and maintains relations among independent and dependent variables. It is the most vital and impressive method being used in data analysis in mining of data. Regression Analysis helps us to assume the mean value of the variable that is dependent provided the value of variable that is independent. It is a mathematical way of sorting out which of the independent or dependent variable does really make an impact. It incorporates numerous techniques for demonstrating and examining a several variables, when the attention is on the relationship between one independent variables and dependent variable [8]. Regression analysis is often used in statistics to obtain data trends. For instance we can develop a connection between the amount of food we eat and the weight we gain. This tern "Regression" was initially used by Francis Galton in early nineties to explain a biological phenomenon. Regression analysis is processing the data using various data processing abilities and statistical algorithms to develop patterns and mutual relations in existing data base [9]. It is shown in Figure 1.

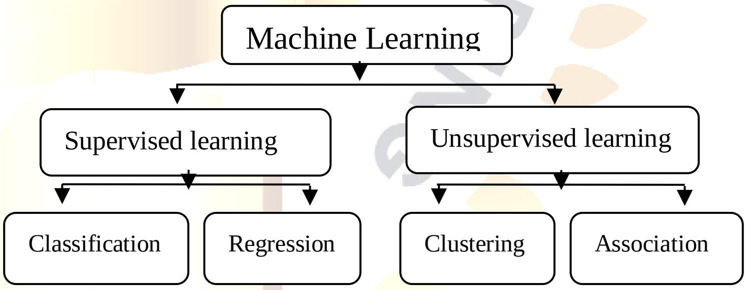

Figure 1: Regression Analysis and Classification

3. Data Mining

Data Mining is a method of extracting valuable information that is useful from the vast pool of data stored in data warehouses which is complex and unstructured to arrive at some decisions. Data mining involves various methods ranging from data discovery, data processing, data analyzing and mining to find patterns and relationships inside data.

Data mining is thus a confluence of various other frontiers or fields like statis- tics, artificial intelligence, machine learning, database management, pattern recognition, and data visualization as pre- sented in Figure 2.

In simple words, **data mining is** defined as a process used to extract usable **data** from a larger set of any raw **data**.

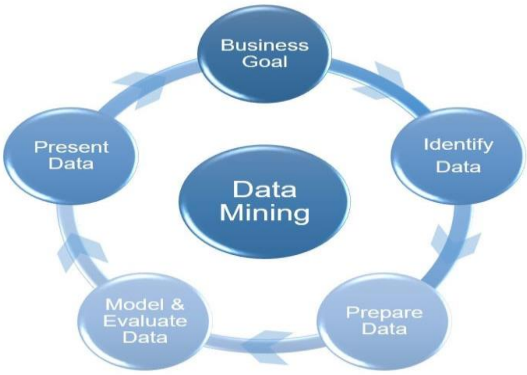

Figure 2: Data Mining

4. LITERATURE REVIEW

4.1 Multiple Regression Analysis of Performance Indicators

Turóczy *et al.* [13] presented the research methodology which is based on statistical analysis, this paper includes the multiple regression analysis. This type of analysis is applied for modelling or analysing on a few variables. The multiple regression analysis expands regression analysis Titan et al., by describing the relationship between a dependent or a single independent variables Constant. It analysis the simultaneous concern that some independent variables have more than one dependent variable; it can be utilized for predicting and forecasting. The multiple regression models can be considerably more real compared to unifactorial regression. In our study the dependent variable consists in the profit size, while the independent variables are the following: self- financing capacity, return on value, personnel cost per employee and investment per person employed. These variables were observed all through ten years. To begin with we presented the essential data for the analysis, after which we obtained the regression equation. We calculated the coefficient of determination R_2, which had the point of indicating the precent of the amount of the aggregate change is clarified by the independent variables.

4.2 Privacy Preserving

Aggarwal *et al.* [14] presented randomization of exact values using Gaussian and Uniform perturbations. Their algorithm dependent on a Bayesian procedure for improving perturbed distributions. Then, Verykios *et al.* [15] categorized the present privacy preserving algorithms in five various categories: Data or rule hiding, privacy preservation, data distribution, data modification and DM algorithm.

Bertino *et al.* [16] categorized present PPDM algorithms from a proposed taxonomy. Ketel *et al.* [17] introduced a geometric rotation based data perturbation. Privacy pre- serving classification methods avoid a miner from classifier construction which can predict sensitive data. Privacy pre- serving clustering techniques that interfere with sensitive numerical attributes, although preserving general features have been proposed.

Another works on random projection or random rotation of hybridization methods by Ramu *et al.* [18]. The hybrids have been tested on four bankruptcy or six benchmark problems. Logistic regression, Decision tree and MLP have been applied for categorizing using 10-fold cross-validation. Bansal *et al.* [19] introduced a novel algorithm because preserving privacy through neural network learning. Ravi *et al.* [20] introduced a novel privacy preservation technique namely particle swarm optimization performing Auto- Associative Neural Network

4.3 Data Mining Techniques

The prediction, is one of the data mining techniques that determines relationship among dependent variables or in- dependent variables. The prediction analysis technique be used into sale to predict profit, Here sale is an indepen- dent variable, and profit may be a dependent variable. It is based on the past sale and profit data, a regression curve that is used for profit prediction. The difficulty in prediction a data is a complex set [21]. In fact there are no methodologies or tools can ensure to generate the exact prediction in the organization. In this paper, they have examined the distinctive algorithm and prediction procedure. Inspite the way that the least median squares regression is known to produce better results than the classifier linear regression techniques from the given set of attributes. As correlation they found that Linear Regression method which takes the lesser time when contrasted with Least Median Square Regression [22]. The data analytics approaches can be followed to predict the target customer which can be of different types for example Statistical Analytics or Dynamic Analytics [23].

5. PROPOSED APPROACH

In this paper we propose Regression Modelling Technique which manages the correlation and association between statistical variables. The variables here are treated symmetrically.

5.1 Regression Analysis

The data can be analyzed with the help of statistical ana- lytic technique. These techniques include Linear Regression, which is the simplest form of regression. It models a random variable, *Y* called a response variable, as a linear function of another variable X which is called as a Predictor Variable. Thus the equation becomes according to linear Regression is:

$$y = a + bX$$

Where the variance of Y is assumed to be constant, a and b are regression coefficients which specifies the Y-intercept and slope of line. The coefficients can be solved with the method of Least Squares, which helps in minimization of the data between the actual data and the estimated line.

5.2 Multiple Linear Regression

In this section, we review briefly the multiple regression model that you encountered in the DMD course. There is a continuous random variable called the dependent variable, Y , and a number of independent variables, x_1, x_2,...,x_p. Our purpose is to predict the value of the dependent variable (also referred to as the response variable) using a linear function of the independent variables. The values of the in- dependent variables (also referred to as predictor variables, repressors or covariates) are known quantities for purposes of prediction, the model is:

$$Y = \beta_0+\beta_1x_1+\beta_2x_2+ \ +\beta_px_p+s,$$

Where s, the "noise" variable, is a normally distributed random variable with mean equal to zero and standard deviation σ whose value we do not know. We also do not know the values of the coefficients β_0, β_1, β_2, , β_p. We estimate all these (p + 2) unknown values from the available data. The data consist of n rows of observations also called cases, which give us values y_i, x_{i1}, x_{i2}, . . . , x_{ip}; i = 1, 2, , n. The estimates for the β coefficients are computed so as to minimize the sum of squares of differences between the fitted (predicted) values at the observed values in the data.

6. Conclusion and Future Scope

The paper presents a distributed data mining approach, suitable for modeling and prediction of numerical quantities. The approach is based on optimization over input data composed exclusively of numerical attributes. The numerical data are frequently used in data mining in fields such as chemistry, physics, and hydrology. The simplicity of such produced model (it takes the form of a regression function), and its usefulness in the reasoning phase are among its most prominent advantages. The mechanism of choice of model structure also proved to be very useful; on one hand, it gives the possibility to choose the universal structure of the regress function, on the other hand, it allows forcing a function specially designed for a particular case within the process.

References

[1] B. Buelens, P. Daas, and J. van den Brakel, "Data min- ing for official statistics: Challenges and opportunities," in *2012 IEEE 12th International Conference on Data Mining Workshops*, pp. 915–915, Dec 2012.

[2] X. Wu, X. Zhu, G. Q. Wu, and W. Ding, "Data mining with big data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, pp. 97–107, Jan 2014.

[3] D. Che, M. Safran, and Z. Peng, *From Big Data to Big Data Mining: Challenges, Issues, and Opportunities*, pp. 1–15. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013

[4] N. E. Oweis, S. S. Owais, W. George, M. G. Suliman, and V. Snášel, *A Survey on Big Data, Mining: (Tools, Techniques, Applications and Notable Uses)*, pp. 109–119. Cham: Springer International Publishing, 2015.

[5] L. Shan and Z. Xuefeng, "The application of data mining in statistics of r amp;amp;d," in *2012 International Conference on Computer Science and Service System Aug 12*

[6] P. Yang, X. Gui, F. Tian, J. Yao, and J. Lin, "A privacy-preserving data obfuscation scheme used in data statistics and data mining," in *2013 IEEE 10th International Con- ference on High Performance Computing and Communica- tions 2013 IEEE International Conference on Embedded and Ubiquitous Computing*, pp. 881–887, Nov 2013.

[7] L. Igual and S. Seguı́, *Regression Analysis*, pp. 97–114. Cham: Springer International Publishing, 2017.

[8] K. Adachi, *Regression Analysis*, pp. 47–62. Singapore: Springer Singapore, 2016.

[9] P. V. Jirapure and P. A. Deshkar, "Qualitative data analy- sis using regression method for agricultural data," in *2016 World Conference on Futuristic Trends in Research and In- novation for Social Welfare (Startup Conclave)*, pp. 1–6, Feb 2016.

[10] V. Ribeiro, A. Rocha, R. Peixoto, F. Portela, and M. F. San- tos, "Importance of statistics for data mining and data science", in 2017

[11] B. Buelens, P. Daas, and J. van den Brakel, "Data min- ing for official statistics: Challenges and opportunities," in *2012 IEEE 12th International Conference on Data Mining Workshops*, pp. 915–915, Dec 2012.

[12] P. Yang, X. Gui, F. Tian, J. Yao, and J. Lin, "A privacy-preserving data obfuscation scheme used in data statistics and data mining," in *2013 IEEE 10th International Conference on High Performance Computing and Communications 2013 IEEE International Conference on Embedded and Ubiquitous Computing*, pp. 881–887, Nov 2013.

[13] Z. Turóczy and L. Marian, "Multiple regression analysis of performance indicators in the ceramic industry," *Procedia Economics and Finance*, vol. 3, no. Supplement C, pp. 509 – 514, 2012. International Conference Emerging Markets Queries in Finance and Business, Petru Maior University of Tˆırgu-Mures, ROMANIA, October 24th - 27th, 2012.

[14] C. C. Aggarwal and P. S. Yu, *An Introduction to Privacy-Preserving Data Mining*, pp. 1–9. Boston, MA: Springer US, 2008.

[15] V. S. Verykios and A. Gkoulalas-Divanis, *A Survey of Association Rule Hiding Methods for Privacy*, pp. 267–289. Boston, MA: Springer US, 2008.

[16] E. Bertino, I. N. Fovino, and L. P. Provenza, "A framework for evaluating privacy preserving data mining algorithms*," *Data Mining and Knowledge Discovery*, vol. 11, pp. 121–154, Sep 2005.

[17] M. Ketel and A. Homaifar, "Privacy-preserving mining by rotational data transformation," in *Proceedings of the 43rd Annual Southeast Regional Conference - Volume 1*, ACM-SE 43, pp. 233–236, 2005.

[18] K. Ramu and V. Ravi, "Privacy preservation in data min- ing using hybrid perturbation methods: an application to bankruptcy prediction in banks," *International Journal of Data Analysis Technique and Strategies* vol 1 no 4.

[19] A. Bansal, T. Chen, and S. Zhong, "Privacy preserving back-propagation neural network learning over arbitrarily parti- tioned data," *Neural Computing and Applications*, vol. 20,

[20] Paramjeet, V. Ravi, N. Naveen, and C. R. Rao, "Privacy preserving data mining using particle swarm optimisation trained

auto-associative neural network: an application to bankruptcy prediction in banks," *International Journal of Data Mining, Modelling and Management*, vol. 4, no. 1

[21] N. M. M. Ramos, J. M. P. Q. Delgado, R. M. S. F. Almeida, M. L. Simões, and S. Manuel, *Data Mining Techniques*,

[22] H. A. Madni, Z. Anwar, and M. A. Shah, "Data min- ing techniques and applications, a decade review," in *2017 23rd International Conference on Automation and Comput-*

[23] M. Rathi, *Regression Modeling Technique on Data Mining for Prediction of CRM*, pp. 195–200. Berlin, Heidelberg: Springer Berlin 2010