Review of Regression Analysis in Data Mining

Afreen Ali [1], Sarwesh Site [2],
[1,2]*Computer Science & Engineering Department,*
All Saint's College of Technology, Bhopal, India

***Abstract*—Regression analysis is a statistical processes which is widely used in data mining and big data analysis. In this paper, we are surveying the various regression analysis techniques for analyzing data for Big Data and Data Mining. There are various proposed regression modeling method for the retention of data analysis in the business aspects. The aim of this paper is how the concept of data mining or regression analysis can be used on Big Data and real data sets with positive results. Main focus of univariate regression is analyze the relationship between dependent and independent variable and conveys the linear relation equation between independent and dependent variable.**

***Keywords*— Data Mining, Regression Analysis, Big Data, Statistical Analysis.**

I. Introduction

The activity of 'Regression Analysis' is to make quantitative expectations of one variable from the estimations of another variable [1]. Today Big data are collected in many areas, and modern database systems allow effective storing and fast access on Big data. Also, increasing performance of the computers and allows the execution of complex processing in very short time [2][3]. These are very effective conditions for the use of data mining and information recovery in many domains of industry and science. Big data are taking a lead, which poses a number of problems, their processing is time, memory demanding,and requires complex parallelization strategies. In mining with big data, it is possible to choose a representative training set of appropriate size, remaining data will be included into testing set [4]. Data manipulation is so much easier and faster. Data mining tools predict future patterns roles, enabling businesses to make proactive, knowledge driven decisions. The automatic, prospective analysis provided by data mining move beyond the analysis of past events allowed by review tools typical of decision support systems [5][6].

II. Regression Analysis

Regression analysis [7] is a statistical tool for evaluating the connections among variables. Regression analysis is most valuable technique used in data mining. Regression analysis is a statistical tool for the examination of relationship between variables. 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 a data mining function that predicts a number, age, weight, distance, sales, income and temperature, could all be predicted utilizing regression techniques. A regression task start with a data set in which the target values are known the objective of regression analysis is to complete the parameters estimations for a function that generate the function to best fit a collection of observed data that provide. Regression analysis is processing the data using complex data search capabilities and statistical algorithms to search patterns and correlations in large prior databases [9]. It is represented in Figure 1.

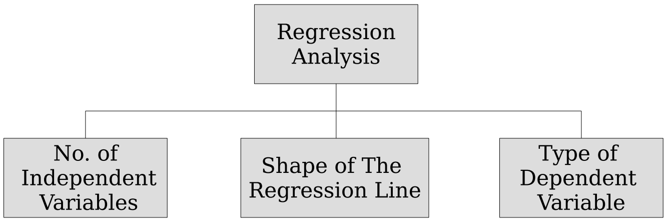
Regression Analysis
No. of Independent Variables
Shape of The Regression Line
Type of Dependent Variable

Fig. 1. Regression Analysis and Classification

III. Data Mining

Data Mining is the process for extracting previously unknown, logical, and actionable data from large databases and applying it to start on crucial business decisions. Data Mining is a set of methods utilized as a part of the knowledge discovery process to distinguish previously unknown relationships and patterns inside data. Data mining is thus a confluence of various other frontiers or fields like statistics, artificial intelligence, machine learning, database management, pattern recognition, and data visualization as presented in Figure 2.

Data mining is the "automated extraction of unknown predictive information from databases".

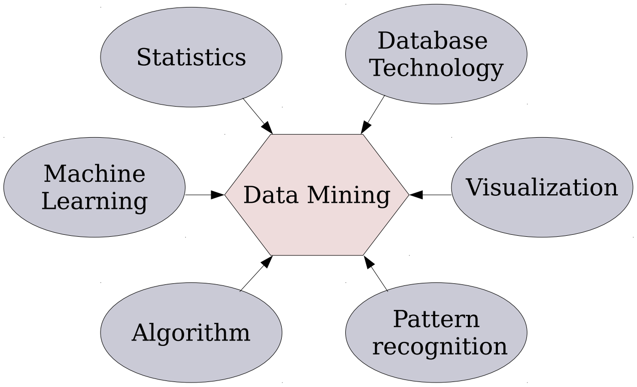

Fig. 2. Frontiers of Data Mining

Statistics in Data Mining

Statistics is a component of data mining that provides the tools and analytics techniques for dealing with large amount of data. It is the science of learning from data or includes everything from collecting and organizing to analyzing and presenting data [10]. It is concerned with probabilistic models, specifically inference, using data. While the aims of statistics and data mining are comparable, it is evaluated that there are very few analysts to manage the requests of data analysts [11][12].

IV. Literature Review

A. Multiple Regression Analysis of Performance Indicators in Ceramic Industry

Turczy *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 applyed for modeling or analyzing 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 model can be considerably more real compared to unifactorial regression.

Zhang *et al.* [4] proposed a ridge regression approach to big data. Ridge regression is important and has been extensively used in applications. The classical ridge regression approach focuses on small or moderate data. It assumes that the entire data set can be loaded to the memory of a personal computer. However, if the data set is large, then it cannot be loaded to memory, which means that the classical approach cannot be used. To solve the problem, they propose new methods and algorithms, where the entire data set is only scanned once. The goal of scanning data is to compute a matrix of sufficient statistics, which is not large. Once the matrix of sufficient statistics is derived, all of the rest computations can be completely carried out without the need of the original data set. Therefore, our numerical algorithms are efficient.

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 percent of the amount of the aggregate change is clarified by the independent variables. Than we turned to $\mathcal{F}-$test and to Student test, respectively t with $n-(k+1)$ degrees of freedom, in order to see which hypothesis can be accepted.

B. Privacy Preserving Data Mining

Aggarwal *et al.* [14] presented randomization of exact values using Gaussian and Uniform perturbations. Theirs 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] catogrized present PPDM algorithms from a proposed taxonomy. Ketel *et al.* [17] introduced a geometric rotation based data perturbation. Privacy preserving classification methods avoid a miner from classifier construction which can predict sensitive data. Privacy preserving 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 applyed 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. The work deal with classification problems. Logistic regression and Decision tree have been used for data mining purpose.

C. Data Mining Techniques

The prediction, is one of the data mining techniques that determines relationship among dependent variables or independent variables. The prediction analysis technique be used into sale to predict profit, Here sale is an independent 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].

D. Regression Modeling Technique on Data Mining for Prediction of CRM

For implementing best CRM different success factors are available. The data analytics approaches can be followed to predict the target customer which can be different types for example Statistical Analytics or Dynamic Analytics [23]. Further for the data Analysis there can be two forms that can be applied on extract models which describes important data classes and predict future data directions. The Prediction can be done on the basis of historical data which predicts the uncertainty of data sets that whether the customer will be satisfied by the product or not. There are two different types of values discrete type values or continuous valued attributes.

Here our proposed work will be to predict the continuous values using regression techniques.

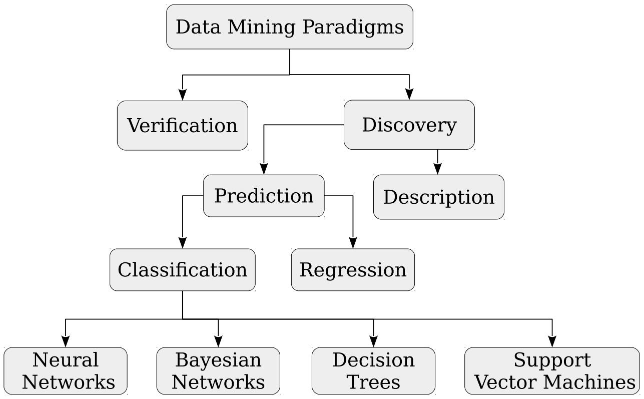

Fig. 3. Prediction in Data Mining

V. Proposed Approach

In this paper we proposed Regression Modeling Technique which manages the correlation and association between statistical variables, the variables here are treated in a symmetrically.

A. *Statistical Analysis*

This section utilizes mainly two statistical methods for data analysis - factor analysis and regression analysis. Factor analysis is used to ensure that the intended constructs can be justified, and to prevent that variables that do not represent what they were intended to measure are included in the final model. After the constructs have been developed, regression analysis is used to test the established hypotheses. These two statistical methods will be briefly described.

1) Factor Analysis: Factor analysis attempts to analyze underlying variables, or factors that shows the pattern of correlations within a set of variables observations. Factor analysis is including data reduction to identify a small number of factors that explain most of the variance that is observed in a much larger number of variables. The purpose of data reduction is to remove redundant (or highly correlated) variables from the data.

In this analysis, principal component analysis was used, which is similar, more reliable and conceptually less complex than “traditional” factor analysis. Principal component analysis is concerned with establishing what kind of linear components that exist in the data and how each variable might contribute to that component. For simplicity, principal component analysis will just be called factor analysis as both methods are very similar.

2) Regression Analysis: The data can be analyzed with the help of statistical analytic 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.
where,

$$\text{Slope}(b) = \frac{n\sum xy - \left(\sum x\right)\left(\sum y\right)}{n\left(\sum x^2\right) - \left(\sum x\right)^2}, \qquad (1)$$

$$\text{Intercept}(a) = \frac{\sum y - b\sum x}{n} \qquad (2)$$

B. *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 independent variables(also referred to as predictor variables, regressors or covariates) are known quantities for purposes of prediction, the model is: $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p + \epsilon$, where ϵ, the noise variable, is a Normally distributed random variable with mean equal to zero and standard deviation whose value we do not know. We also do not know the values of the coefficients $\beta_0, \beta_1, \beta_2, ..., \beta_p$. 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. The sum of squared differences is given by

$$\sum_{i=1}^{n}(y_i - \beta_0 - \beta_{1xi1} - \beta_{2xi2} - ... - \beta_{pxip})^2$$

Let us denote the values of the coefficients that minimize this expression by
$\hat{\beta}_0$,$\hat{\beta}_1$,$\hat{\beta}_2$,...,$\hat{\beta}_p$.These are our estimates for the unknown values and are called OLS (ordinary least squares) estimates in the literature. Once we have computed the estimates $\hat{\beta}_0$,$\hat{\beta}_1$,$\hat{\beta}_2$,...,$\hat{\beta}_p$ we can calculate an unbiased estimate $\hat{\sigma}^2$ for σ^2 using the formula: $\hat{\sigma}^2$= $\dfrac{1}{n-p-1}$

$$\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \beta_{1xi1} - \beta_{2xi2} - ... - \beta_{pxip})^2$$

We plug in the values of $\hat{\beta}_0$,$\hat{\beta}_1$,$\hat{\beta}_2$,...,$\hat{\beta}_p$ in the linear regression model (1) to predict the value of the dependent value from known values of the independent values,$x1, x2, ..., xp$. The predicted value, $\hat{Y}$, is computed from the equation $\hat{Y}$ =$\hat{\beta}_0$ + $\hat{\beta}_{1x1}$+ $\hat{\beta}_{1x2}$ + ... + $\hat{\beta}_{pxp}$.

VI. Conclusion and Future Work

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 mining for official statistics: Challenges and opportunities,” in *2012 IEEE 12th International Conference on Data Mining Workshops*, Dec 2012, pp. 915–915.

[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, no. 1, pp. 97–107, Jan 2014.

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

[4] T. Zhang and B. Yang, “An exact approach to ridge regression for big data,” *Computational Statistics*, vol. 32, no. 3, pp. 909–928, Sep 2017.

[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 2012, pp. 1631–1634.

[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 Conference on High Performance Computing and Communications 2013 IEEE International Conference on Embedded and Ubiquitous Computing*, Nov 2013, pp. 881–887.

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

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

[9] P. V. Jirapure and P. A. Deshkar, “Qualitative data analysis using regression method for agricultural data,” in *2016 World Conference on Futuristic Trends in Research and Innovation for Social Welfare (Startup Conclave)*, Feb 2016, pp. 1–6.

[10] V. Ribeiro, A. Rocha, R. Peixoto, F. Portela, and M. F. Santos, “Importance of statistics for data mining and data science,” in *2017 5th International Conference on Future Internet of Things and Cloud Workshops (FiCloudW)*, Aug 2017, pp. 156–163.

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

[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*, Nov 2013, pp. 881–887.

[13] Z. Turczy 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 Trgu-Mures, ROMANIA, October 24th - 27th, 2012.

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

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

[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, no. 2, 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*, ser. ACM-SE 43, 2005, pp. 233–236.

[18] K. Ramu and V. Ravi, “Privacy preservation in data mining using hybrid perturbation methods: an application to bankruptcy prediction in banks,” *International Journal of Data Analysis Techniques and Strategies*, vol. 1, no. 4, pp. 313–331, 2009.

[19] A. Bansal, T. Chen, and S. Zhong, “Privacy preserving back-propagation neural network learning over arbitrarily partitioned data,” *Neural Computing and Applications*, vol. 20, no. 1, pp. 143–150, Feb 2011.

[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, pp. 39–56, 2012, pMID: 45135.

[21] N. M. M. Ramos, J. M. P. Q. Delgado, R. M. S. F. Almeida, M. L. Simões, and S. Manuel, *Data Mining Techniques*. Cham: Springer International Publishing, 2016, pp. 13–30.

[22] H. A. Madni, Z. Anwar, and M. A. Shah, “Data mining techniques and applications, a decade review,” in *2017 23rd International Conference on Automation and Computing (ICAC)*, Sept 2017, pp. 1–7.

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