A Comparative Analysis on Linear Regression and Support Vector Regression

Kavitha S
Assistant Professor
Computer Science and Engineering
Bannari Amman Institute of Technolgy
Sathyamangalam
kvth.sgm@gmail.com

Varuna S
Assistant Professor
Computer Science and Engineering
Bannari Amman Institute of Technolgy
Sathyamangalam

Ramya R
Assistant Professor
Computer Science and Engineering
Bannari Amman Institute of Technolgy
Sathyamangalam

***Abstract*— In business, consumers interest, behavior, product profits are the insights required to predict the future of business with the current data or historical data. These insights can be generated with the statistical techniques for the purpose of forecasting. The statistical techniques can be evaluated for the predictive model based on the requirements of the data. The prediction and forecasting are done widely with time series data. Most of the applications such as weather forecasting, finance and stock market combine historical data with the current streaming data for better accuracy. However the time series data is analyzed with regression models. In this paper, linear regression and support vector regression model is compared using the training data set in order to use the correct model for better prediction and accuracy.**

***Index Terms*—regression, linear regression, support vector, prediction, data analytics**

I. Introduction

Analytics focus on inference by statistical and mathematical analysis of data. The analysis helps to identify the problem from the collected data source. The solutions or other decisions can be provided with the data analytics tool like Online Analytical Processing (OLAP). Later it uses various tools and algorithms for better outcomes of data. Data Analytics [1] can be descriptive, predictive and perspective. Descriptive analytics summarizes the data. Predictive analytics helps in predicting future outcomes whereas perspective analytics includes predictive and feedback system to track the outcomes produced by action. There are many technologies used in the data analytics but predictive analytics is the one that uses machine learning algorithms and statistical analysis for future prediction. The predictive analytics is used with the emerged as business intelligence applications.

A. Business Intelligence

Business Intelligence [4] is the latest technology comprises with data analytics that drives the predictive data market. The technology provides tools and other software solutions for gaining the business insights to rule over the competitive business market. These applications help the business people to make better decisions. BI applications also help to identify the current trend of the market and to address the business problems. These applications include cleansing data. The data sources for this analysis are obtained from the dash boards, reviews, blogs and discussion forum. These current streaming data, historical data and consumer insights are combined to predict the future events.

B. Machine Learning

Machine learning combines computer science and statistical analysis to enhance the prediction. High value predictions can be obtained within human computer interaction. It helps to predict uncertain situation with the data. Machine learning algorithms are categorized as supervised and unsupervised. Supervised algorithms are used to learn the labeled data and produce the result whereas unsupervised data does not use labeled data for learning. It simply used to obtain the inference from data source. Here supervised learning algorithms of regression model are used for analyzing time series data.

In this paper we analyze linear and support vector regression models in order to use the correct model for prediction based on its requirements. This paper further consists of Related work in Section II, Supervised Machine Learning in Section III, Regression Analysis in Section IV, Experimental Settings in Section V and Conclusion and Future Work in Section VI.

II. RELATED WORK

The data analytics has grown in tremendous after the evolution of Big data technologies. The data analytics with big data technologies results with some data mining tools for big data analytics [1], [2]. The major purpose of these analytics is prediction which is said to be Business Intelligence [3], [4]. Predictive analytics [5] rules the business market with BI applications. These analytics are performed with machine learning algorithms [6] over the data. There are several machine learning algorithms used for classification and regression. Mainly supervised learning models [7] are preferred because of its training data model. The supervised learning algorithms for linear regression [8], [9] have several regression functions. Linear regression with LeastMedSq function is used to find the minimum square of all median and In support vector regression there are several functions subjected to linear and non-linear kernels. Support Vector Regression [10], [11], [12], [13] with linear kernel can be used for linear regression. In SVR, Sequential Minimal Optimization (SMO) regression [14] function is used along with the linear kernel function [15], [16]

978-1-5090-4556-3/16/$31.00 ©2016 IEEE

for linear regression. The models are evaluated with the metrics such as Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) to estimate its performance to use the model for particular business applications.

III. SUPERVISED MACHINE LEARNING

Supervised Machine Learning represented in Fig 1 infers a function from labeled training data. The training data is mapped to new value from the input data and produce result [7]. This method is faster and produces accurate results. The training dataset consist number of tuples (T). Each tuple is a vector which contains attribute values. The target data can have more than one possible outcome or a continuous value. $(T \cup X)$ denotes T as input attributes that contains n attributes. T = {$a_1, a_2, a_3, \ldots\ldots a_n$} and y as target attribute. The attributes can be unordered set values or real numbers. The training sets are assumed to be generated randomly. The supervised learning algorithm can be used for classification and regression.

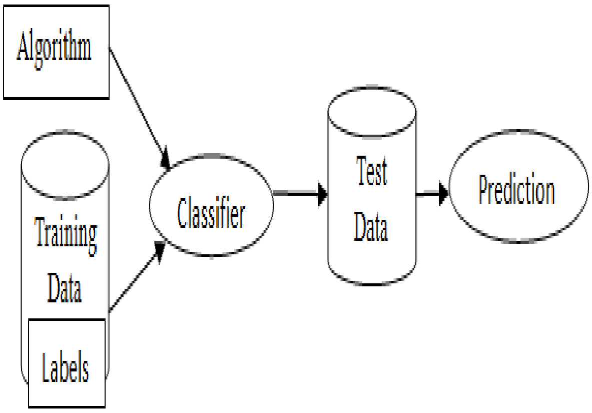

Fig. 1. Supervised Machine Learning

A. Classification

Classification is the process of predicting labels based on categories represented in Fig 2. For example we can classify marks of a student as pass or fail. For classification process, the classifier is built and it is used for further classification of data. Building classifier is the learning process. It can be built using the classification algorithm from the training data set. Each tuple consists of training data set with the class label. Then the test data is used to evaluate the classification rules using the classifier. The training attributes can have the values of nominal and numeric whereas the target attribute can have more than two possible outcomes.

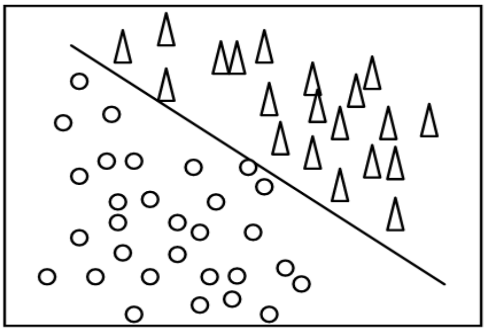

Fig. 2. Classification

B. Regression

Regression is a statistical analysis method to identify the relationship between the variables. The relationship can be identified between the dependent and independent variables. It can be described using probability distribution functions represented in Eq.1.

$$Y = f(X, \beta) \tag{1}$$

Where Y is a dependent variable, X is an independent variable and β is an unknown parameter.

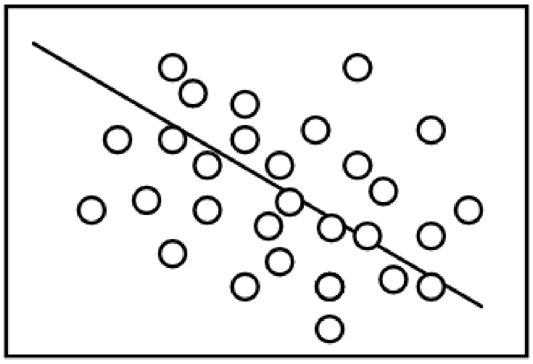

Fig. 3. Regression

The variable dependency can be either univariate or multivariate regression. Univariate regression identifies the dependency among single variable as represented in Eq. 2.

$$y = a + bx + \varepsilon \tag{2}$$

Where y is a dependent variable, x is an independent variable with co-efficient b and a is a constant. while multivariate regression [17] is to identify the dependency among several variables simultaneously is represented in Eq. 3.

$$y = a + b_1x + b_2x + \ldots. + b_nx + \varepsilon \tag{3}$$

In this paper, multivariate analysis is done with both regression models.

IV. REGRESSION MODELS

Regression models predict the outcome of the dependent variables from the independent variables. The significance is considered in regression analysis to handle the most complicated problems. In this paper, we discuss about linear and support vector regression which best fits the predictive model.

A. Linear Regression

Linear Regression [8], [9] is the most common predictive model to identify the relationship among the variables. Apart from univariate or multivariate data types the concept is linear. Linear regression can be either simple linear or multiple linear regression. The linear regression is described in Eq. 4.

$$y = x\beta + \varepsilon \tag{4}$$

In Eq. 4. y is the independent variable which can be either continuous or categorical value, x is a dependent variable which is always a continuous value. It is analyzed with probability distribution and mainly focused on conditional probability distribution with multivariate analysis.

1) Simple Linear Regression

Simple linear regression represented in Fig. 4 is the process of prediction using single independent variable which is univariate regression analysis as described in Eq.2. Simple linear regression distinct the dependent variables and independent variables to extent the relationship between two variables as similar to correlation but correlation does not distinct the dependent and independent variables.

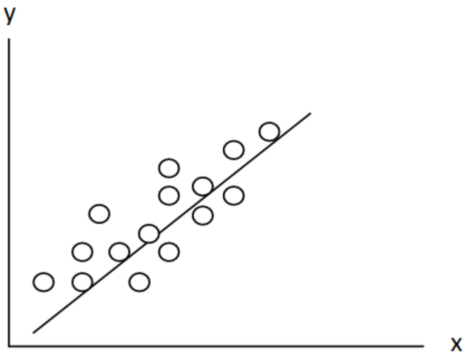

Fig. 4. Simple Linear Regression

2) Multiple Linear Regression

Multiple or multi variable linear regression represented in Fig. 5 is the process of prediction with more than one independent or predictor variables which is similar to multivariate analysis as described in Eq. 3.

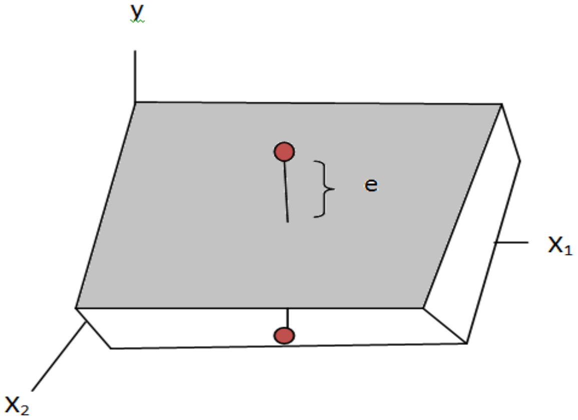

Fig. 5. Multiple Linear Regression

a) LeastMedSq Linear Regression

In this regression model, random samples are used to create the least median square functions. The model is evaluated with the value obtained for the median square which should be very minimum for its best fit.

B. Support Vector Regression

Support Vector Machine is one of the supervised learning models for classification and regression [12], [13]. Support Vector Machine for regression is specifically said to be Support Vector Regression. Support Vector Regression can be linear or non-linear using respective kernel functions.

1) *Linear Support Vector Regression*

Support Vector Regression uses linear kernel functions for regression which is similar to support vector machines but SVR sets the tolerance margin (ε) to approximation not like SVM which should be taken from the problem is represented in Fig, 6.

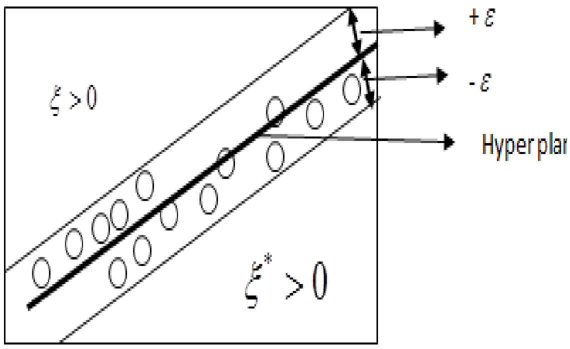

Fig. 6. Support Vector Linear Regression

Support Vector Regression with linear kernel function is described in Eq. 5.

$$y = w.x + b \tag{5}$$

Where the input space is denoted by y, vector product is denoted as w.x and b is a constant. Based on the error function Eq. 6, it can be minimized where the target will be $z_{i.}$

$$\min \frac{1}{2}\left\|w^2\right\| + c\sum_{i=1}^{n}\left(\xi_i + \xi_i^*\right) \tag{6}$$

Subjected to

$$\begin{cases} z_i - (w.x + b) & \le \varepsilon + \xi_i \\ (w..x) + b - z_i & \le \varepsilon + \xi_i^* \\ \xi_i \xi_i^* & \ge 0 \end{cases}$$

a) SMOReg Linear Kernel

SMOReg implements Support Vector Regression with various kernels. In this paper we use SMOReg with linear kernel function for analysis of linear regression. Linear Kernel function is described in Eq. 7.

$$K(x, y) =< x, y > \tag{7}$$

2) Non Linear Support Vector Regression

In Non-Linear Support Vector Regression, the non-linear kernel functions are used for processing the training data in feature space as represented in Fig. 7.

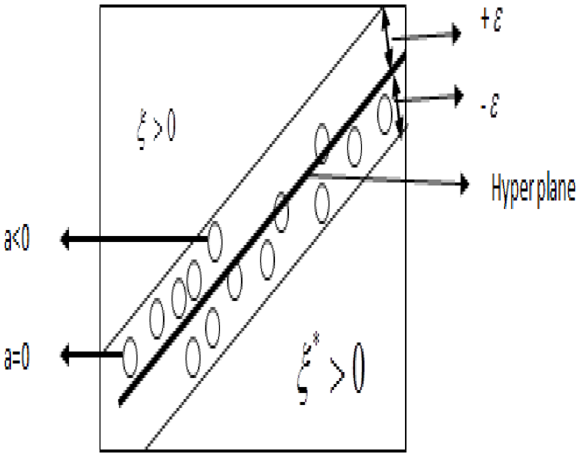

Fig. 7. Non-Linear Support Vector Regression

After processing the training data into feature space, then normal support vector regression is applied using Eq. 8.

$$\max\begin{cases}\frac{1}{2}\sum_{i,j=1}^{n}(\alpha_i-\alpha_i^*)(\alpha_j-\alpha_j^*k)(x_{i,}x_j)\\ -\varepsilon\sum_{i=1}^{n}(\alpha_i-\alpha_i^*)+\sum_{i=1}^{n}y_i(\alpha_i-\alpha_i^*)\end{cases} \quad (8)$$

Subjected to

$$\sum_{i=1}^{n}(\alpha_i-\alpha_i^*)=0; \qquad 0\le\alpha_{i,}\alpha_i^*\le C$$

Where $\alpha_{i,}\alpha_i^*$ are Lagrange multipliers

V. Experimental Settings

This section shows the experimental setup and results of linear regression models evaluated using training data set for the best fit of the model that can be used for regression analysis of time series data

A. Experimental Setup

In this paper, the experiment analysis was carried out using an existing data analysis tool called Weka [18], [19]. The dataset has been collected from the public data repository called UCI data repository [20]. The assessment data of students is taken for multivariate analysis. The dataset consists of 1000 instances which in turn consist of 5 attributes. The details of the dataset is represented in Table-I. The evaluation is based on training set. The dataset will be classified automatically as train and test data sets.

TABLE I. Data Definition

Data Set	Open University Learning Analytics Dataset
Data Set Characteristics	Multivariate, Sequential, Time Series
No. of Instances	1000
No. of Attributes	5
Attributes	id_assessment id_student date_submitted is_banked score
Attribute Characteristics	Integer
Test Mode	Evaluate on training set
Classifier Model	Full Training set

B. Experimental Results

Table II and Table III represents the metrics obtained for the evaluation of LeastMedSq function for linear regression model and SMOreg function which uses SVM with linear kernel model respectively.

TABLE II. LeastMedSq Fsunction(Training set)

METRICS	OBSERVED VALUE
Time taken to build model	3.29 seconds
Correlation coefficient	-0.0038
Mean absolute error	9.6689
Root mean squared error	12.54
Relative absolute error	98.8263%
Root relative squared error	100.9408%
Total Number of Instances	1000

The evaluation metrics in the above table specifies only linear regression model with LeastMed sq function.

TABLE III. SMOreg Function (Training set)

METRICS	OBSERVED VALUE
Time taken to build model	2.42 seconds
Correlation coefficient	-0.0029
Mean absolute error	10.0784
Root mean squared error	12.8725
Relative absolute error	103.0124%
Root relative squared error	103.6172%
Total Number of Instances	1000

The evaluation metrics in the above table specifies only linear regression model with SMOreg function with linear kernel where c=1.0.

C. Performance Evaluation

The performance is evaluated with the metrics obtained. Here in this analysis we consider only two metrics i.e., Mean absolute error and Root mean squared error for evaluation.

1) Evaluation Metrics

a) Mean absolute error

It calculates the mean of all absolute errors for all predicted values which is described in Eq. 8.

$$\sqrt{\frac{\sum_{i=1}^{n} y_i - y_i'}{n}} \qquad (8)$$

b) Root mean squared error

It calculates the square root of all mean squared errors which is described in Eq. 9.

$$\frac{\sum_{i=1}^{n} \left| y_i - y_i' \right|}{n} \qquad (9)$$

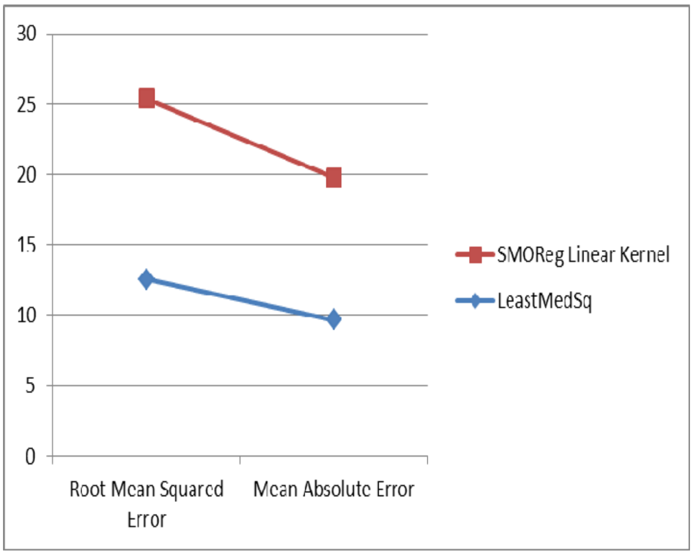

Fig. 8. Comparison of LinearMedSq and SMOReg Linear Kernel functions

It is observed that LinearMedSq function best fits for linear regression though the time taken for building the training model is more compared to SMOReg function with linear kernel.

VI. Conclusion and Future Work

Data Analytics and business intelligence plays a major role in the current competitive market. In case of analyzing a time series multivariate analysis, efficient data model should be used for accurate results. If the linear regression model is used then there are several functions associated with this model. In this paper we analyzed the linear regression model with LeastMedSq function and SMOreg function over a multivariate and time series data set. The analytical results concluded that LeaseMedSq is the best model for linear regression.

In future, the better model can be identified for linear regression with minimized RMSE and MAE and also minimum time taken for constructing model on training data.

References

[1] Philip Russom, Big Data Analytics, TDWI Best Practices Report, 2011.

[2] Alfredo Cuzzocrea, Il-Yeol Song,Karen, C. Davis, "Analytics over Large-Scale Multidimensional Data: The Big Data Revolution!", DOLAP'11, ACM, October 28, 2011.

[3] James R. Evans, Carl H. Lindner, "Business Analytics: The Next Frontier for Decision Sciences", Decision Science Institute, March 2012.

[4] Surajit Chaudhuri, Umeshwar Dayal, and Vivek Narasayya, "An Overview of Business Intelligence Technology", Communications of the ACM, Vol. 54, No. 81.1, August 2011.

[5] Galit Shmueli, Otto R. Koppius, "Predictive Analytics in Information Systems Research", Mis Quarterly, Vol. 35, No. 3, pp. 553-572, September 2011.

[6] RS Michalski, JG Carbonell, TM Mitchell, "Machine learning: An artificial intelligence approach",Springer-Verlag, 2013.

[7] S. B. Kotsiantis, "Supervised Machine Learning: A Review of Classification Techniques", J. of Informatica, Vol. 31, pp.249-268, 2007.

[8] GAF Seber, AJ Lee, "Linear regression analysis", Wiley Series in Probability and Statistics,2012.

[9] DC Montgomery, EA Peck, GG Vining, "Introduction to linear regression analysis", Wiley Series in Probability and Statistics, 2015.

[10] Zhang Xuegong, "Introduction to Statistical Learning Theory and Support Vector Machines", Acta Automatica Sinica, 2000.

[11] Smits G.F., Jordaan E.M., "Improved SVM Regression Using Mixtures of Kernels", IJCNN '02. Proceedings of the International Joint Conference on Neural Networks, Vol.3, 2002.

[12] Alex J. Smola and Bernhard Scholkopf," A tutorial on support vector regression", Statistics and computing, Springer, 2004.

[13] SR Gunn, "Support Vector Machines for Classification and Regression", ISIS technical report, 1998.

[14] C Li, L Jiang, "Using locally weighted learning to improve SMOreg for regression", Trends in Artificial Intelligence, PRICAI 2006.

[15] http://crsouza.blogspot.in/2010/03/kernel-functions-for-machine-learning.html

[16] http://crsouza.blogspot.in/2010/04/kernel-support-vector-machines-for.html

[17] Hair, J.F.," Multivariate data analysis", Upper Saddle River, NJ [etc.]: Pearson Prentice Hall, 2006.

[18] Mark Hall, "The Weka Data Mining Software: An Update", ACM SIGKDD Explorations, Vol. 11, No. 1, pp. 10-18, June 2009.

[19] Remco R. Bouckaert, "Weka Manual for Version 3-7-8", The University of Waikato, January 2013.

[20] UCI Machine Repository, http://archieve.ics.uci.edu/ml/datasets/Open+University+/Learning+Analytical+dataset