Saudi Journal of Engineering and Technology (SJEAT)
Scholars Middle East Publishers
Dubai, United Arab Emirates
Website: http://scholarsmepub.com/

ISSN 2415-6272 (Print)
ISSN 2415-6264 (Online)

Prediction Model with Selection of Best Prediction Algorithm for Big Data

S. Banumathi[1*], A. Aloysius[2]
[1]Assistant Professor, Department of Computer Science, Holy Cross College, Trichy, Tamil Nadu, India
[2]Assistant Professor, Department of Computer Science, St. Joseph's College, Trichy, Tamil Nadu, India

Original Research Article

***Corresponding author**
S. Banumathi

Article History
Received: 26.02.2018
Accepted: 07.03.2018
Published: 30.03.2018

DOI:
10.21276/sjeat.2018.3.3.9

Abstract: Big data can be defined as high volume, high velocity and high variety of data that require high performance processing. Predictive modelling is a process of creating a model to predict the future behavior from the data. A predictive model is made up of predictors which are factors that influence future results. A Proposed prediction model is identifies best prediction algorithm by evaluating the performance of an algorithm. The model contains preprocessing module and predicting module which trailed by performance evaluation to find best prediction algorithm. The preprocessing module emphasizes flow of preprocessing methods. The prediction module combines various efficient algorithms to prediction, amid them the best prediction algorithm can be identified by performance evaluation measurement.

Keywords: Big data, Prediction model, Prediction Algorithm, Data Preprocessing Methods.

INTRODUCTION

In the recent domain we incredulous with dealing with petabytes of data. The presence of big data offers both an opportunity as well as challenges to analyses. The term big data was created to define the collection of large amounts of data in structured, semi-structured, or unstructured formats in large databases, file systems, or other types of repositories, and the processing of this data in order to produce an analysis and synthesis of the trends and actions in real or almost real-time[4]. Out of the above amounts of data, the unstructured data needs more real-time analysis and bears more valuable information to be discovered, providing a more in-depth understanding of the researched subject.

It is also the unstructured data which incurs more challenges in collecting, storing, organizing, classifying, analyzing, as well as managing [2].

A proportion of development has been made in progressing the capability to process, store, and analyze big data. In addition to the big data computing capability, the rapid advances in using intelligent data analytics techniques drawn from the emerging areas of artificial intelligence (AI) and machine learning (ML) which provide the ability to process massive amounts of diverse unstructured data that is now being generated daily to extract valuable actionable knowledge[3, 7]. There are four types of analytics namely prescriptive, predictive, diagnostic, and descriptive.

PREDICTIVE ANALYTICS

Predictive analytics refers to a technology that learns from experience to predict the future behavior of individuals in order to drive better decisions. The need to devise a new tool for predictive analytics for both types increases rapidly. Predictive analytics encompasses data science, machine learning, predictive and statistical modeling and outputs empirical predictions based on given input empirical data. The underlying premise is that future can be predicted on the basis of the past experience. The predictive analytics involves data collection, data analysis, data modeling and visualization of data. The analytics has three types of models Predictive model, Descriptive model and Decision model. Predictive modelling is a process of creating a statistical or mathematical model to predict the future behavior from the data set. A predictive model is made up of predictors which are factors that influence future results [6]. The selection of prediction variable influence the prediction accuracy in the model selection [8]. The Descriptive model quantifies the relationships between data to classify into groups. The decision model involves predicting decision by comprise all the data variables and find relationship between data elements.

PREDICTION MODEL

Prediction model involves the process of creating, testing and validating a model to best predict the probability of an outcome. The model creation consists of one or more algorithms for the data set. After creation, the model is tested and validated for

evaluating the best prediction model which fitted for the data set selected. The process of model involves running of more than one algorithm on data set to find predictions [5]. The model process is an iterative process which can create multiple models and finally arriving the best fit model. This paper proposed a new prediction model Fig. 1 to find best prediction algorithm. This model consists of preprocessing module and predictive module, evaluation module. The outcome from the evaluating module results the best prediction algorithm.

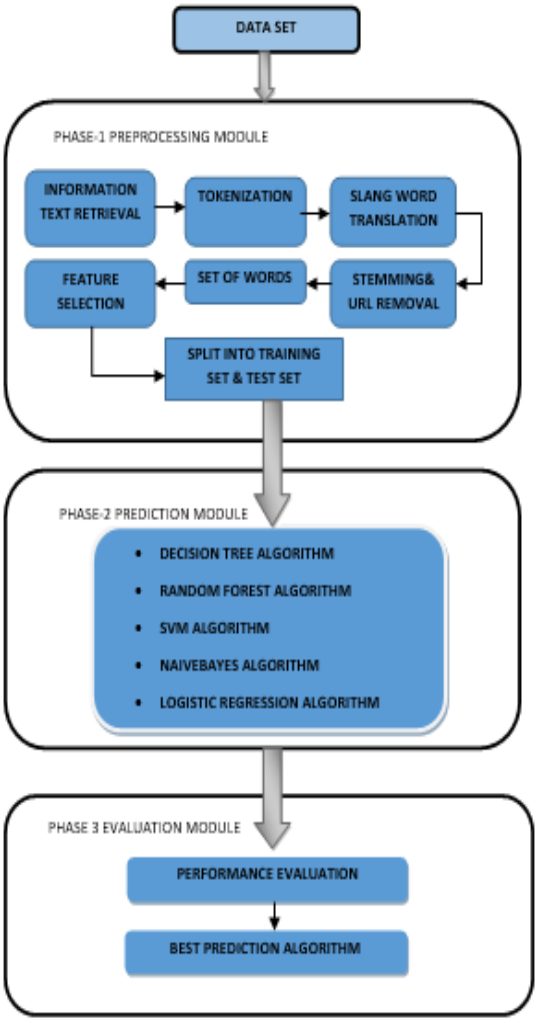

Fig-1: Prediction Model

PREPROCESSING MODULE

Data preprocessing methods can be divided into instance based methods and regression based methods. The Preprocessing methods are used to eliminate unwanted things in the data set. Different preprocessing techniques are used to make the data for testing and training. The information is retrieved in the form of unstructured and preprocessing methods is applied to make data suitable for further analysis [9]. Tokenization is the process of extracting bags of cleaner terms from raw comments by deleting stop words and punctuation, compressing redundant character repetitions and deleting IDs or name used in the text for messaging purposes. For removing stop word, stop word dictionary maintained and used which contain all stop words. Each word of comments is compared with this dictionary and the matched word gets removed from comment. The User generated comments often contains the slang words. Slang word translation means converting the slang words like lol, omg etc, into their standard form. The Internet Slang Word Dictionary is used for this and then add them to the comments. Stemming means a group of different words share the same meaning. It is the process of reducing words which share the same meaning. We used stem word dictionary for grouping all different words of same meaning. In URL removal URLs in the data set is removed in which many users add URL. These URLs make the process of applying prediction algorithm more complex. So the URL's are removed. Then the feature selection algorithm is applied to select the features for making the data set into training and testing.

PREDICTION MODULE

In this proposed model the prediction module includes various evolutionary algorithms which used for prediction. Each algorithm has their own features as pros and cons. The tool has been used to apply an algorithm to data. History cannot always predict future: Using relations derived from historical data to predict the future implicitly assumes there are certain steady-state conditions or constants in the complex system. This is almost always wrong when the system involves people. The issue of unknown unknowns: In all data collection, the collector first defines the set of variables for which data is collected. However, no matter how extensive the collector considers his selection of the variables, there is always the possibility of new variables that have not been considered or even defined, yet are critical to the outcome.

EVALUATION MODULE

The proposed evaluation module considers performance evaluation of various algorithms used in prediction module. Performance evaluation usually measured as the terms of mean square error and average relative error. This module evaluates the performance analysis based on the error values. The comparison of error measurement values will identify the best prediction algorithm for given data.

CONCLUSION

In this paper a new prediction model proposed with four phases. Each phase get the input from behind phase. They illustrated as modules, each module processes and as the overall frame work output will be the best prediction algorithm for big data. This increase the credibility of an algorithm in prediction. The future work will experiment the framework with streaming data as input and identify the prediction algorithm which will best suitable to data.

REFERENCES

1. García, S., Ramírez-Gallego, S., Luengo, J., Benítez, J. M., & Herrera, F. (2016). Big data preprocessing: methods and prospects. *Big Data Analytics*, *1*(1), 9.
2. Ionescu, B., Ionescu, D., Gadea, C., Solomon, B., & Trifan, M. (2016). An Architecture and Methods for Big Data Analysis. In *Soft Computing Applications* (pp. 491-514). Springer, Cham.
3. Khan, N., Husain, M. S., & Beg, M. R. (2015). Big data classification using evolutionary techniques: a survey. In *Proceedings of IEEE International Conference on Engineering and Technology (ICETECH)* (pp. 243-247).
4. Sagiroglu, S., & Sinanc, D. (2013, May). Big data: A review. In *Collaboration Technologies and Systems (CTS), 2013 International Conference on* (pp. 42-47). IEEE.
5. Jung, S. H., Kim, J. C., & Sim, C. B. (2016). Prediction data processing scheme using an artificial neural network and data clustering for Big Data. *International Journal of Electrical and Computer Engineering*, *6*(1), 330.
6. Anderson, C. W., Lee, M., & Elliott, D. L. (2015, July). Faster reinforcement learning after pretraining deep networks to predict state dynamics. In *Neural Networks (IJCNN), 2015 International Joint Conference on* (pp. 1-7). IEEE.
7. Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*, *35*(2), 137-144.
8. Huang, S., & Fang, N. (2013). Predicting student academic performance in an engineering dynamics course: A comparison of four types of predictive mathematical models. *Computers & Education*, *61*, 133-145.
9. Babu, U. R. (2017). Sentiment Analysis of Reviews for E-Shopping Websites. *International Journal of Engineering And Computer Science*, *6*(1).