OPEN ACCESS

International
Journal
Of Modern Engineering Research (IJMER)

Predictive Analytics Using Decision Tree In Big Data

D.Praveena[1], Dr.M.SureshKumar[2]

[1]*Department of Computer Science & Engineering, PG Scholar, Sri Ramakrishna Engineering College, Coimbatore, India*

[2]*Department of Computer Science & Engineering, Professor, Sri Ramakrishna Engineering College, Coimbatore, India*

ABSTRACT : *Predictive analysis and models are typically used to forecast future probabilities. They are used to analyze the current data and historical facts to identify various potential risks and opportunities and to understand the system better. Big data is enabled by Predictive analytics approach. Businesses collect vast amounts of real-time customer data and predictive analytics uses this historical and aperiodic data, with the combination of customer insight, in order to predict future events. Predictive analytics redirects organizations to utilize big data both real-time and stored in order to migrate from a historical view of processing data to a customers' forward-looking perspective. Predictive analytics is used to foretell any behavior on analysis. It could also be applied to any scenario where customers' or machine behavior needs to be predicted. An analysis and prediction framework is proposed. The framework accepts large number of observations as an input. These observations are trained and a model is created. The test data is evaluated by analyzing, predicting and forecasting the future behavior of the user provided historic data. The performance evaluation of the proposed matching algorithm clearly defines that it is a better method for prediction and analysis*

Keywords: *Cluster, Hadoop, Precision, Predictive analytics, Supervised algorithm.*

I. INTRODUCTION

Big data is a collection of huge data sets which is very hard to process using traditional data processing applications because of its larger size and complexity[1]. The larger data sets is because of the extra information which is derivable from the analysis of a single large set of relevant data, when compared to separate smaller sets with the same total amount of data [2]. Data sets grow in size in a rapid manner since they are increasingly collected by the ubiquitous information-sensing mobile devices, software logs, aerial sensory technologies, microphones, cameras, wireless sensor networks and radio-frequency identification readers [3]. Usually Big data includes datasets with large sizes beyond the ability of most commonly used software tools to capture, process and handle the data within tolerable elapsed time [3]. The data size is rapidly increasing from terabytes to many petabytes and even zettabytes of data on one single data set.

A) NEED FOR BIG DATA

[1] Raise of storage capacities
[2] Requirement for high processing power
[3] Availability of data
[4] Better business decisions including both strategic and operational

B) ISSUES IN BIG DATA

In many cases, how different data sets are related to each other is not known [4]. So it can be found through a process of exploration and discovery.The Actual relationship are not always known in advance, so the uncovering insight is an iterative process [5]. Many industry experts and analysts suggest to start with tiny, well-defined projects, study from every single iteration, and gradually migrate towards the next field of inquiry or idea [6]. The discovery process includes data exploration to understand how it can be used but also determines how well it is related to the traditional enterprise data [7].

II. PREDICTIVE ANALYTICS USING BIG DATA

Big data trend always leaves lot of room for malfunctions, if data is not analyzed thoroughly. In existing system, one or two low incidences of high value critical observations go undetected. They use statistical models.

Suspicious claims are identified and then analyzed by special investigation units, claim adjuster etc.,[8]. Sampling methods are used for analysis which leads to one or more observations going undetected. Existing services have new, unforeseen risk factors [8]. Most existing solutions are inflexible and not built for today's line of business. They rely on previously existing cases, so every time when a new case occurs they have to bear the consequences. They are incapable to handle the ever growing information sources from various channels and different functions in an integrated way [9]. They use outdated supervised algorithms for the prediction of observations.

[5] *ARCHITECTURE DIAGRAM*

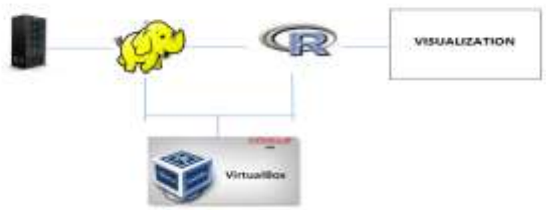

Figure 1: Architecture diagram for predictive analytics

Fig.1 describes the architecture of the proposed system. It is based on the control flow between the loaded data, Hadoop framework and R. The framework makes use of a multi node cluster (i.e. both Master and Slave resides on the different nodes). The data input is used to create statistics and prediction about the risk. The statistical data are framed and submitted to the master which makes use of the Hadoop framework to process the data stored in the HDFS and displays the retrieved results to the local file system.

The complete end to end architecture is divided into the following major tasks: Statistical analysis; Distributed processing; Predictive analytics; Visualization. Working directory is being created on loading the data in the IDE and all the required analysis tasks are being done.These analytical tasks are distributed across various nodes of the Hadoop cluster to perform the required task as the process of parallel processing [9].The statistical and prediction results which are processed in the multiple nodes are collectively consolidated in the master node of the Hadoop cluster [10] - [13]. These predicted results are then visualised in the master node using various plots.

III. PROCESS WORKFLOW

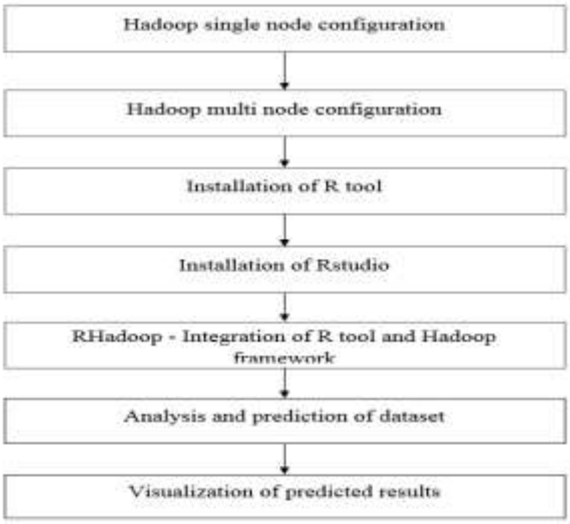
Hadoop single node configuration
Hadoop multi node configuration
Installation of R tool
Installation of Rstudio
RHadoop - Integration of R tool and Hadoop framework
Analysis and prediction of dataset
Visualization of predicted results

Figure 2: Implementation work flow

Fig.2 explains that once when Hadoop single and multi-node are configured, R tool is installed on top of Hadoop distributed framework. R, an open-source data mining tool is used with an integrated development environment named Rstudio. R code is written for analysis of the data set since Hadoop is basically developed from java. The code consists of attributes of the data set used for analysis. Master distributes the job to various slave nodes and the resulting statistics and prediction results are saved as a CSV file in the master node after consolidation.

A) ADVANTAGES

[6] High scalability and precision.
[7] Improves the performance.
[8] Optimizes the processing time consumption.
[9] Ease of visualization

Public survival prediction data set is used. After Hadoop single and multi-node are configured, R tool is installed on top of Hadoop distributed framework. R code is written for analysis of the data set. The code consists of attributes of the data set used for analysis. In order to run these R jobs parallely across multiple nodes, the job of R and Hadoop need to be integrated. This is done by the installation of RHadoop [14]. The jobs are distributed by the master node across various slave nodes and the result displays the entire statistics and prediction results which is saved as a CSV file in the master node. The R code is run in Rstudio. Supervised learning algorithm namely decision tree is used to build the model for the test set evaluation. After analysis of the data set, the results can be directly viewed in the local file system according to the user needs. Results are stored in the local file system. They are gathered and depicted in the form of charts using visualization tools.

IV. RESULTS AND DISCUSSION

The performance of the analysis is compared with many different supervised algorithms in order to measure the efficiency of the system. For this purpose, data set is classified in to test and trained set to analyze the outcomes of the proposed matching algorithms. The performance of processing is evaluated using metrics such as precision and recall. For the data set, the precision and recall values of the algorithms are computed. Precision is defined as the condition, quality or fact of being exact and accurate. Recall is the relevancy fraction of the data which is appropriate to analysis that are retrieved successfully. The recall and precision values are given in Table.1.

Table 1: Precision - Recall Evaluation Result

CLASSIFIER MODEL	TRAINED	TEST	BUILD TIME	PRECISION	RECALL
RPART	93.52%	91.12%	<20 sec	0.8246000	0.7987503
CART	90.01%	86.79%	<90sec	0.6602317	0.7667543
C 5.0	92.51%	91.08%	<60 sec	0.3838384	1.0000000
CHAID	92.51%	89.37%	<60 sec	0.6402314	0.0043875
ADABOOST	90.21%	90.05%	<50 sec	0.7454700	0.7309942

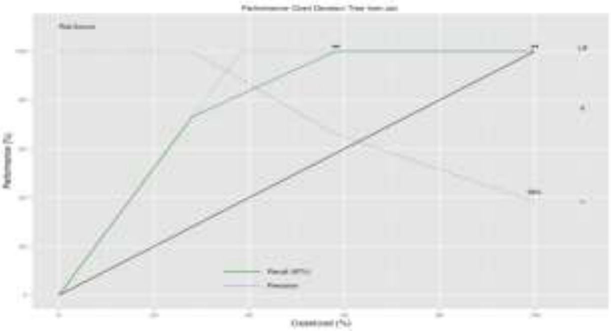

Figure 3: Precision - Recall Graph

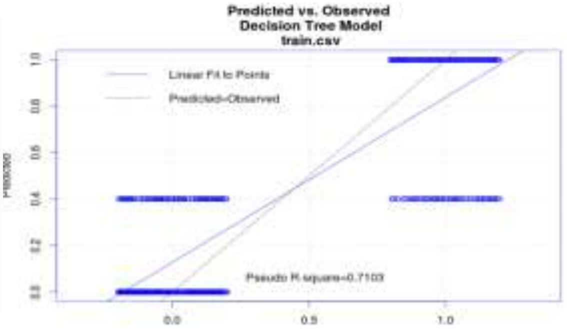

Figure 4: Predicted vs. Observed decision tree model

Fig.3 illustrates the comparison between the precision (blue curve) and recall (green curve). The precision values are plotted in X-axis and the corresponding recall values for the proposed system are plotted in Y-axis respectively. It is obvious that the proposed work yields better results than the other existing algorithms. Fig.4 illustrates the comparison between the predicted result and survived value is compared. The survived values are plotted in X-axis and the corresponding predicted values for the proposed system are plotted in Y-axis.

V. Conclusion

Since users often have little knowledge about supervised algorithms and implementation details, a framework which performs analysis and prediction on voluminous and variety of data is needed. The proposed framework presents a prediction mechanism that enables analysis and prediction on big data and its associated concepts and the future work is aimed at extending the approach to use other data mining algorithms to build new classification models on any of the data set in real world and the performance of the new models will be compared with the existing models. Also, rather than comparing performance alone over the prediction accuracy the future comparisons will be extended in a manner to include these comparisons over other performance metrics to make the analysis much more efficient.

References

[1]. Amlan Kundu, Suvasini Panigrahi, Shamik Sural, Arun K. Majumdar, (2009) "Credit card fraud detection: A fusion approach using Dempster–Shafer theory and Bayesian learning," Special Issue on Information Fusion in Computer Security, Vol. 10, Issue no 4, pp.354- 363.

[2]. Ularu, E. G., Puican, F. C., Apostu, A., & Velicanu, M, (2012) "Perspectives on Big Data and Big Data Analytics".

[3]. Ana-Ramona Bologa, Razvan Bologa, Alexandra Florea, (2010) "Big Data and specific analysis methods for Insurance fraud detection".

[4]. William Hendrix et al., (2011) "Community Dynamics and Analysis of Decadal Trends in Climate Data".

[5]. Anuj Sharma, Prabin Kumar Panigrahi, (2012) "A Review Of financial accounting fraud detection based on data mining techniques".

[6]. Kaiyong Deng, Ru Zhang, Hong Guo,Kaiyong Deng,R Zhang, Dongfang Zhang, WenFeng Jiang, Xinxin Niu (2011)"Analysis and Study on Detection of Credit Fraud in E-commerce".

[7]. A.Shen, R.Tong, and Y.Deng, (2007) "Application of classification models on credit card fraud detection".

[8]. Khyati Chaudhary, Bhawna Mallick, (2012) "Credit Card Fraud: Bang in E-Commerce".

[9]. Abhinav Srivastava, Amlan Kundu, Shamik Sural, Arun K. Majumdar, (2008) "Credit Card Fraud Detection using Hidden Markov Model," IEEE Transactions On Dependable And Secure Computing, vol. 5, Issue no. 1, pp.37-48.

[10]. Md Delwar Hussain Mahdi, Karim Mohammed Rezaul, Muhammad Azizur Rahman, (2010) "Credit Fraud Detection in the Banking Sector in UK: A Focus on E-Business".

[11]. Venkata Reddy Konasani, Mukul Biswas, Praveen Krishnan Keloth, (2012) "Health care fraud management using big data analytics".

[12]. M.F.Gadi, X. Wang, and A.P. Lago, (2008) "Comparison with parametric optimization in credit card fraud detection".

[13]. Y. Sahin, E. Duman, (2011) "Detecting Credit Card Fraud by ANN and Logistic Regression".

[14]. Phua, C., Lee, V., Smith, K., & Gayler, R, (2010) "A comprehensive survey of data mining-based fraud detection research".