International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-8 Issue-12, October 2019

Tobit Regressive Based Gaussian Independence Bayes Map Reduce Classifier on Data Warehouse for Predictive Analytics

R. Sivakkolundu, V. Kavitha

Abstract- Data warehouse comprises of data collected from different probable heterogeneous resources at different time intervals with the objective of responding to user analytic queries. Big data is a field that helps in analysing and extracting information from large datasets. The unfolding Big Data incorporation inflicts multiple confronts, compromising the feasible business research practice. Heterogeneous resources, high dimensionality and massive volumes that confront Big Data prototype may prevent the effectual data and system integration processes. In this work, we plan to develop a Tobit Regressive based Gaussian Independence Bayes Map Reduce Classifier (TR-GIBMRC) method for categorizing the collected and stored data which helps the users in making decision with minimum time consumption. The TR-GIBMRC method consists of two processes. They are, Tobit Regressive Feature Selection and Gaussian Independence Bayes Map Reduce Classification. Tobit Regressive Feature Selection process is used to select relevant features from collected and stored data. Tobit statistical model, used to describe the relationship between non-negative dependent variable and an independent variable for selecting relevant features. Next, Gaussian Independence Bayes Map Reduce Classifier is used to classify the selected relevant features for decision making with lesser time consumption. Gaussian Independence Bayes Map Reduce Classifier, a probabilistic classifier segments the data by class by measuring the mean and variance of data in each class. The data point gets allocated to the class with minimal variance. This in turn helps to perform efficient data classification for accurate decision making. Experimental evaluation is carried out on the factors such as feature selection rate, classification accuracy, classification time and error rate with respect to number of features and number of data points.

Keywords: Big Data, Data warehouse, Feature Selection, Gaussian, Bayes, Map Reduce Classifier, Tobit Regressive

I. INTRODUCTION

Data warehouse is the place to store information for making decisions. Big data is large volume of structured, semi-structured and unstructured data that has potential to be mined for information. Data warehouse is a way that integrates data from large and inconsistent database located in different locations. Modern data warehouses provided better solutions for radical variations by reducing storage volume via velocity enhancement in enhancing velocity in multidimensional design and through data elaboration for providing qualitative information.

A new framework was introduced in [1] for physical and methodological features of data warehouse by considering factors that affect data warehouse lifecycle. The criteria were set for classifying the Big Data Warehouses due to the methodological features. But volume and veracity problems were not addressed. For Veracity evaluation, data quality model was essential one for categorizing dirty data stored in data warehouse and describing the metrics for counting errors in every class. An evolutionary game theory-based method was introduced in [2] for materialized view selection in data warehouse with multiple view processing plan structure to find problem search space. A population of players were generated where each player were considered as solution to the problem. Three strategies were taken for each player. At each repetition of game, players selected best strategy for themselves. The final solution was determined consistent with strategies chosen by the players. But the designed method failed to reduce classification accuracy using evolutionary game theory-based method. An integrated artifacts were introduced in [3] for resilient multidimensional warehouse repository. The knowledge-based data models were included with spatial-temporal dimensions to reduce ambiguity in warehouse repository execution. The design consideration guaranteed uniqueness and monotonic properties of dimensions, preserving connectivity between artifacts and attaining business alignments. The multidimensional attributes visualized Big Data analyst with valuable knowledge for decision support systems. But the classification time was not reduced for integrated artifacts. In addition, it was not suitable for various applications like healthcare ecosystems, environment modelling and disaster management domains. A state-of-the-art review presenting a holistic view of the challenges involved in Big Data and the methods in Big Data Analytics employed by organizations to assist others in understanding the landscape for investment decisions were presented in [4]. The domain of healthcare has received its effect by the impact of big data due to the fact that the data sources involved in the healthcare organizations are familiar for their volume, heterogeneous entanglement and lofty dynamism. In [5], different analytical avenues that prevail in the patient-centric healthcare system from the angle of several stakeholders were presented. In the last decade, one of the de-facto standard frameworks for processing of big data in several industries is Hadoop. A framework called, Scalding was used in [6] to produce an effective solution to big data image processing with the objective of creating photographic mosaics.

Revised Manuscript Received on October 30, 2019.
***** Correspondence Author
R. Sivakkolundu*, Department of Computer Science, Bharathiar University, Coimbatore, India.
Dr. V. Kavitha, Professor, Department of PG and Research Department of Computer Applications (MCA), Hindusthan College of Arts and Science, Coimbatore, India.

In recent years, several business cases utilizing big data have been realized, to name a few are, Twitter, LinkedIn, and Facebook in the social networking domain. However, conceptual work combining these methods into single reasonable reference architecture has been limited. In [7], technology independent reference architecture for big data systems was designed based on the analysis of implementation architectures of big data use cases. Yet another simplified distributed classifier training model called, Label-aware Distributed Ensemble Learning (LADEL) was designed in [8] to handle Big Data that generated stratified samples that in turn minimized the inter-machine communication time. In the domain of medical field, the volume of data is growing exponentially and hence conventional models cannot manage it in an effective manner. Due to this, the role of big data in the area of management, organization and data analysis utilizing machine and artificial intelligence are growing in a rapid manner. In [9], a system for the administration and scrutiny of biomedical image data based on the tools of big data technology was presented. With the inception of data warehouse, the information pertaining to customers has to be recorded in an intermittent manner. Fuzzy semantics were applied in [10] with the objective of analysing consumer behaviors for accurate customer classification.

Existing methods for predictive analysis on data warehouse towards big data contain limitations that have yet to be resolved. For example, the former cannot perform dimensionality reduction during data processing with lesser time, which may lead to inaccuracies during subsequent classification operations. The latter can only be executed on the foundation of Map Reduce classifier, without which the concept of feature selection and classification cannot be established. Unlike the human brain, computers cannot naturally determine the degree of classification between the collected data by just looking into the features in the collected data. Hence, relevant feature selection with map reduce classifier has yet to be realized, despite advances in modern technology. In this study, the characteristics of collected and stored data are used to construct a new algorithm and analytical process for predictive analysis purposes. First, Tobit Regressive Feature Selection model is used to select the relevant features. Next, a Gaussian Independence Bayes Map Reduce Classifier is built based on the optimal relevant features. When in operation, the proposed system would allow the objective quantification of optimal features and the classification of new analytical variables for decision making. Potential information discovered via this method can provide decision makers in the data ware house management system with favourable marketing strategies that meet customer requirements. The rest of the paper is organized as follows: Section 2 introduces the predictive analysis for big data in the domain of data warehouse and highlights the key pitfalls in the related work. Following that, the Tobit Regressive based Gaussian Independence Bayes Map Reduce Classifier (TR-GIBMRC) method is presented in Section3. Experimental evaluation along with the discussion follows in Section 4. Finally, section 5 has the conclusions.

II. RELATED WORKS

In this section, related work to solve Map Reduce problems with respect to data warehouse is investigated. In today's world, business establishments generate huge sensitive data. Besides, the velocity of digital data increases and overwhelms the storage capability of business establishments. The management of such huge sensitive data is cumbersome to store the digital data locally and hence results in the huge expenditure. In [11], a remote data checking model utilizing Divide and Conquer tables were used for large scale data storage with minimum computational cost. However, the time involved was less concentrated. To address this issue, a method called, Chabok was designed in [12] that utilized Map Reduce model to solve issues related to data warehouse. However, due to limited storage capability, certain data pertaining to call center customers are discarded, therefore compromising accuracy. To address this issue, Hadoop and Mahout were combined in [13] to support technical requests. Target prioritization is receiving increasing interest in the recent years in biomedical research. In [14], a protocol using TargetMine was designed with the objective of identifying known disease associated genes with higher amount of accuracy. A review of big data concerning health research was presented in [15]. Yet another case study for big data science to replicate complex analysis was designed in [16]. A framework for parallelization on big data using Apache Hadoop with its Map Reduce function was investigated in [17]. However, though accuracy was provided in all the above said methods, the security aspect was not covered. To address this issue, in [18], security algorithm interfaced with the data node was provided. Challenges, methodologies and applications related to industrial big data analytics was presented in [19]. Machine learning and Bayesian learning perspectives to big data characterizing data heterogeneity was presented in [20]. Though the above said methods, ensured dimensionality reduction in the perspective of big data, less focus was made on the performance of classification accuracy and classification time during prediction process with data warehouse and big data. In order to improve the classification accuracy and to reduce the classification time, in this work, Tobit Regressive based Gaussian Independence Bayes Map Reduce Classifier (TR-GIBMRC) method is designed. The elaborate description of the TR-GIBMRC method is provided below.

III. TOBIT REGRESSIVE BASED GAUSSIAN INDEPENDENCE BAYES MAP REDUCE CLASSIFIER

In this section, a Tobit Regressive based Gaussian Independence Bayes Map Reduce Classifier (TR-GIBMRC) method is presented to categorize the collected and stored data that assists the users in making decision with minimum time. The design of TR-GIBMRC method involves two steps. They are Tobit Regressive Feature Selection and Gaussian Independence Bayes Map Reduce Classification.

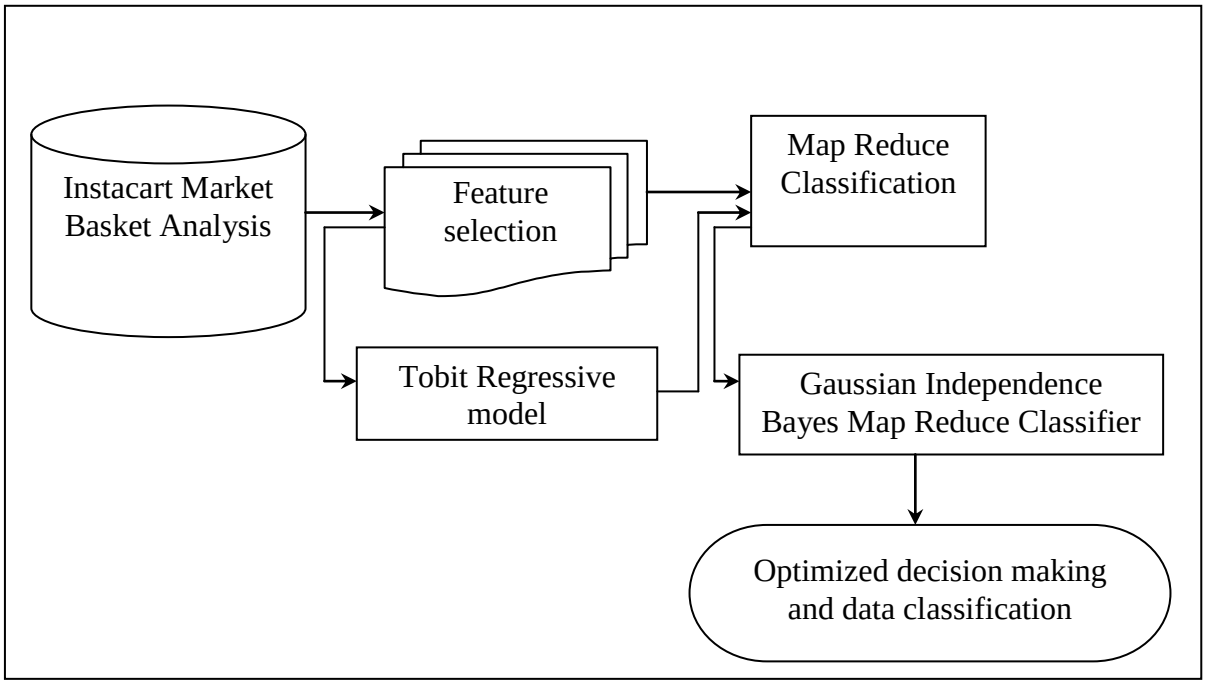

Figure 1 Block diagram of Tobit Regressive based Gaussian Independence Bayes Map Reduce Classifier

As shown in the figure, with the Big Data dataset (i.e., Instacart Market Basket Analysis) from data warehouse, relevant features are selected from the collected and stored data using Tobit Regressive model. The Tobit Regressive model being a statistical model is used in describing the relationship between non-negative dependent variable and an independent variable with the objective of selecting the more relevant features from Big Data dataset. With the relevant feature selection, the collected and stored data are classified using Gaussian Independence Bayes Map Reduce Classifier with help of map reduce function for decision making with lesser time consumption. Gaussian Independence Bayes Map Reduce Classifier being a probabilistic classifier based on Bayes theorem with strong independence assumptions between relevant features classifies in an effective manner according to the Gaussian distribution.

Tobit Regressive Feature Selection

Feature selection is used along with the classifier with the purpose of avoiding over-fitting, to produce more genuine classifier and to dispense more comprehensions into the intrinsic informal relationships. When the input dimension variable is high (i.e., Big Data) compared to the overall sample size, feature selection is customarily required along with a strong classifier. Feature selection helps in providing more discriminations into the elementary causal relationships involved in Big Data by concentrating on lesser number of features, produce more definitive approximates by discarding irrelevant data. In this work, Tobit Regressive Feature Selection model is used to select the relevant features from collected and stored data (i.e. Big Data), i.e. via Instacart Market Basket Analysis. The Tobit model is a statistical model that is utilized to draw the association or correlation between non-negative dependent variable and an independent variable for selecting the relevant features present in Big Data. Figure given below shows the block diagram of Tobit Regressive Feature Selection model.

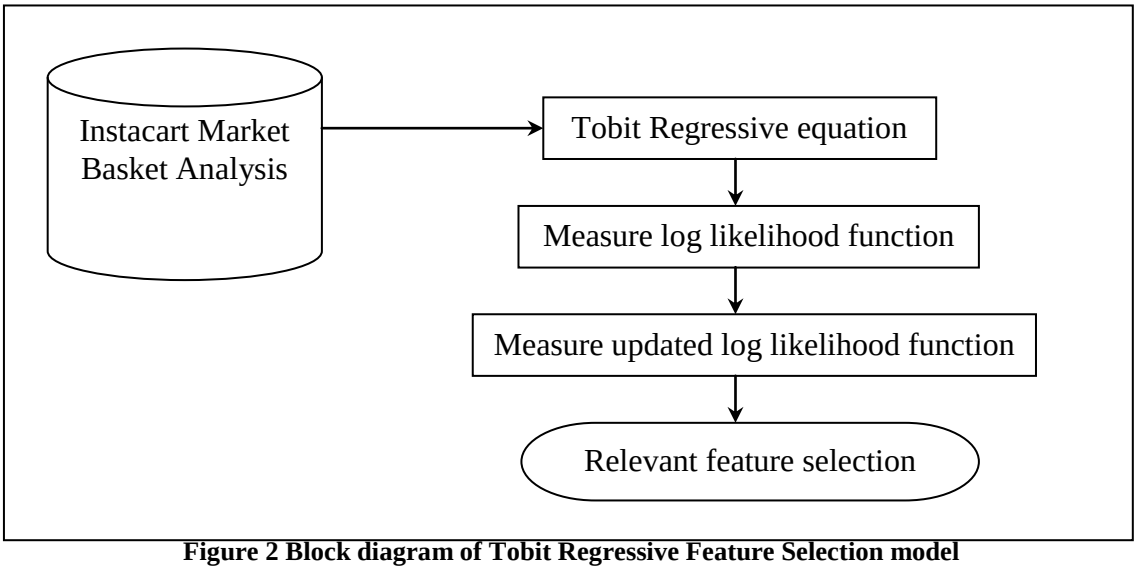

Figure 2 Block diagram of Tobit Regressive Feature Selection model

The Tobit Regressive Feature Selection model is the model in which only distinguished dependent variables are used that satisfies certain constraints. In this model, censoring concept is used to discard the irrelevant features from Big Data. The fundamental equation in the Tobit Regressive model is mathematically formulated as given below.

$$q_i' = P_i\alpha + \epsilon_i \quad (1)$$

From the above equation (1), 'q_i'' is an inactive variable that is noticed for values greater than 'δ' and censored or discarded otherwise. In this work, we are intended in the distribution of 'q' given that 'q' is greater than the cutoff point 'δ' for training (i.e. collected and stored data) 'P_i', with 'ϵ_i' corresponding to the 'N' value lying between '0' and 'σ^2' (i.e., $0.\sigma^2$). In other words, it identifies the relationship between non-negative dependent variable and negative independent variable. This is derived as given below.

$$q_i = \begin{cases} q' \; if \; q' > \delta \\ \delta_q \; if \; q' \leq \delta \end{cases} \quad (2)$$

In the Tobit Regressive model, let us further assume that '$\delta = 0$' i.e. the data are censored or discarded at 0. Thus, the above equation (), is re-written as given below.

$$q_i = \begin{cases} q' \; if \; q' > 0 \\ 0 \; if \; q' \leq 0 \end{cases} \quad (3)$$

Then, the likelihood function for the censored (i.e., irrelevant feature) is mathematically formulated as given below.

$$LF = \left[\frac{1}{\sigma}\left(\frac{q-\mu}{\sigma}\right)\right]^{v_i}\left[\left(\frac{\mu-\delta}{\sigma}\right)\right]^{1-v_i} \quad (4)$$

From the above equation (4), the likelihood function 'LF' is measured using the mean 'μ' of the collected and stored training data 'P_i', variance 'σ' of the collected and stored training data 'P_i' and the index variable 'v'. Here, the index variable 'v' equals to '1', if '$q > \delta$' (here the observation is uncensored or not discarded), and the index variable 'v' equals to '0' if '$q \leq \delta$' (here the observation is censored or discarded). Then, the updated log likelihood function for obtaining censored (i.e. discarded) and uncensored (i.e. not discarded) data is measured as given below.

$$ULF = \left[\frac{1}{\sigma}\left(\frac{q_i-P_i\alpha}{\sigma}\right)\right]^{v_i}\left[\left(\frac{P_i\alpha}{\sigma}\right)\right]^{1-v_i} \quad (5)$$

From the above equation (5), the first part '$\left[\frac{1}{\sigma}\left(\frac{q_i-P_i\alpha}{\sigma}\right)\right]^{v_i}$', corresponds to the uncensored data, whereas the second part '$\left[\left(\frac{P_i\alpha}{\sigma}\right)\right]^{1-v_i}$' corresponds to the censored data. The pseudo code representation of Tobit Regressive Log Likelihood (TRLL) algorithm is given below.

Input: Training data 'P_i', cutoff point 'δ'
Output: optimized and relevant feature selection '$RF = RF_1, RF_2, \dots, RF_n$'
1: Begin 2: For each training data 'P_i' 3: Obtain inactive variable using (1) 4: Measure likelihood function using (2) 5: Measure updated log likelihood function using (4) 6: Obtain censored and uncensored data using (5) 7: Return (optimal and relevant features) 8: End for 9: End

Algorithm 1 Pseudo code of the Tobit Regressive Log Likelihood algorithm

TRLL is the optimal model for solving the relevant feature selection problem in Big Data that is based on censor regressive model. In this model, two different factors are considered for optimal feature selection i.e., censored or discarded data and uncensored or not-discarded data. To identify the relationship between the censored and uncensored data and with the nature of enormous data size involved, Tobit Regression model is considered. In this model, log likelihood function and its updated function based on the index variable is used to identify the non-negative dependent variable and negative independent variable. By identifying this relationship, optimal relevant features are selected, that forms the basis for classification.

Gaussian Independence Bayes Map Reduce Classifier model

With the optimal relevant feature selected, the next step is to classify the features using Gaussian Independence Bayes Map Reduce Classifier with help of map reduce function. With this, the decision-making capacity is said to be improved with minimum time consumption. In this work, a probabilistic classified based on Bayes theorem called, Gaussian Independence Bayes Map Reduce Classifier model is used. This model classifies based on the strong independence assumptions between relevant features by applying the Gaussian distribution. The Gaussian Independence Bayes Map classifier segments the optimal selected features by measuring the mean and variance of data in each class. With this, the data points are said to be allocated to the corresponding class with minimal variance. This in turn ensures in significant data classification that results in precise decision making. Figure 3 shows the block diagram of Gaussian Independence Bayes Map Reduce Classifier model.

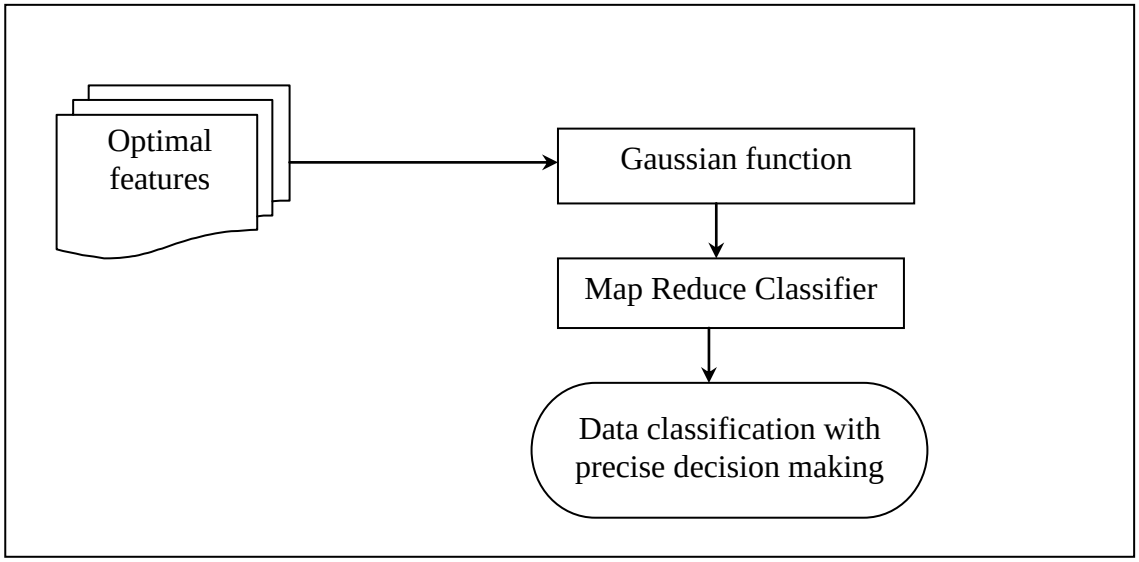

Figure 3 Block diagram of Gaussian Independence Bayes Map Reduce Classifier model

As illustrated in the above figure, the Gaussian Independence Bayes Map Reduce Classifier is one of the machine learning models applied for efficient classification that serves in effective decision making. Let us consider that if there are two classes 'Cl_1' and 'Cl_2' to be classified and each instance has 'N' attributes (i.e., optimal relevant features). In order to estimate the attributes i.e., '$Prob(Cl_j)$', and '$Prob(RF_i = rf|\ Cl_j)$', where '$Prob(Cl_j)$' refers to the prior of class i.e., '$Class\ Cl_j (j = 1; j = 2)$', and '$Prob(RF_i = rf|\ Cl_j)$', corresponds to the likelihood of the 'ith' attribute (relevant feature), 'RF_i' denotes the value '$(i = 1,2,..,N)$' on class 'Cl_j', then, the proposed method measures the total number of instances in the sample

N', occurrences of class 'Cl_j' in the sample, namely 'N_j', and number of instances having 'ith' attribute with value 'rf' in the sample namely 'n_i'.

For the next step in our work, Map Reduce classifier model is used to obtain the value of 'n_i'. In the map phase, user query is taken as input and converts the user query into a '$Key - Value$' pair, where the 'Key' is a combination of the class, attribute and its attribute value, namely, 'RF_i, rf_i, Cl_j'. In the reduce task, the values of the key is appended and a single '$Key - Value$' pair is produced, where the 'Key' is unique strong combination of user query and the value is the frequency of occurrences of such strong user query combination. Figure shows the block diagram of Map Reduce classifier model for Gaussian Independence Bayes.

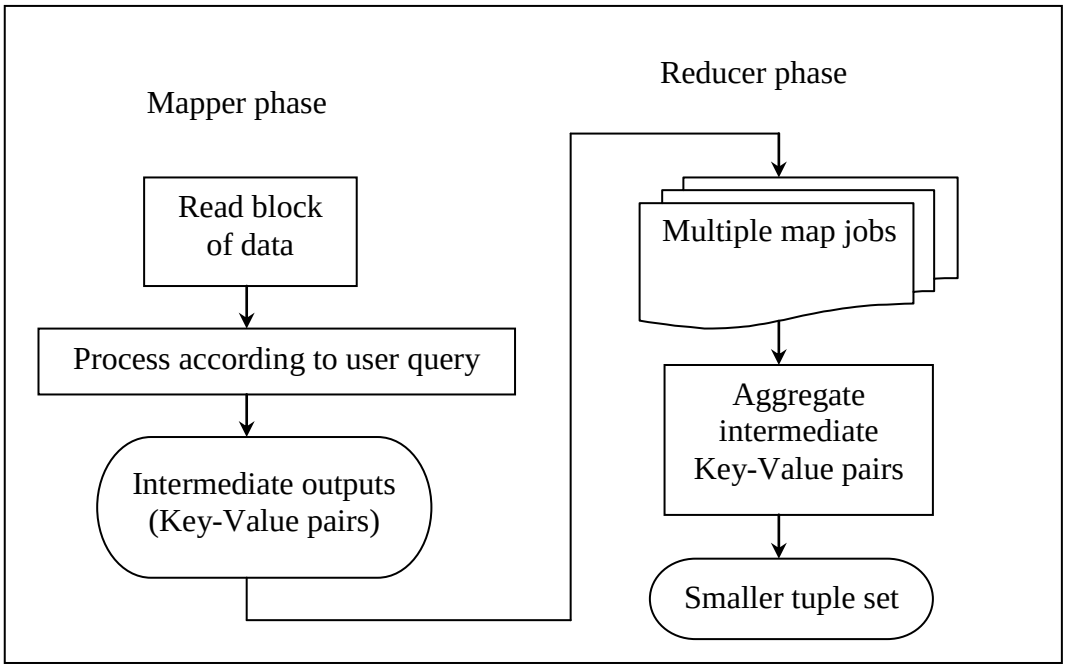

Figure 4 Map Reduce classifier model for Gaussian Independence Bayes

As illustrated in the above figure, with the MapReduce approach, user analytic queries are processed in multiple nodes in a parallel fashion which in turn notably quickens the performance when compared with running on single node. From the resultant MapReduce run, the prior and likelihood for each class is obtained mathematically as given below.

$$Prob(Cl_j) = \frac{N_j}{N} \quad (6)$$

$$Prob(RF_i = rf \mid Cl_j) = \frac{n_i}{N} \quad (7)$$

Besides, from the above equations (6) and (7), the posterior of each class is obtained mathematically as given below.

$$Prob\,(Cl_j|d) * Prob(Cl_j) * Prob(RF_1 = rf_1|\, Cl_j) * Prob(RF_2 = rf_2|\, Cl_j) * \ldots * Prob(RF_N = rf_N|\, Cl_j) \quad (8)$$

The designed classifier given above (8) segments the data by class and then computes the mean and variance of data in each class. This is mathematically formulated as given below.

$$\mu_{ik} = E[RF_i \mid P = p_k] \quad (9)$$

$$\sigma_{ik} = E[(RF_i - \mu_{ik})|\, P = p_k] \quad (10)$$

From the above equations (9) and (10), the mean and variance of data in each class is measured based on each attributed (i.e., related feature) 'RF_i', and each possible value 'p_k' of 'P'. The pseudo code representation of Gaussian Independence Bayes Map Reduce Classifier is given below.

Input: optimal relevant features '$RF = RF_1, RF_2, \ldots, RF_n$', Class '$Cl = Cl_1, Cl_2, \ldots, Cl_n$'
Output: Optimal classification

```
1: Begin
2:     For each optimal relevant features 'RF'
3:         For each class 'Cl'
4:             Obtain prior for each class using (6)
5:             Obtain likelihood for each class using (7)
6:             Measure posterior for each class using (8)
7:             Measure mean for each class using (9)
8:             Measure variance for each using (10)
9:             If 'Prob(Cl1|d) > Prob(Cl2|d)' then
10:                Classify data instance 'd' to class 'Cl1'
11:            End if
12:            If 'Prob(Cl1|d) > Prob(Cl2|d)' then
13:                Classify data instance 'd' to class 'Cl2'
14:            End if
15:        End for
16:    End for
17: End
```

Algorithm 2 Pseudo code of the Gaussian Independence Bayes Map Reduce Classifier algorithm

As given in the above Gaussian Independence Bayes Map Reduce Classifier algorithm for each optimal relevant feature and for each class, three different factors are analyzed. They are prior of each class, likelihood of each class and posterior of each class. Besides, the three factors, the mean and variance of each class are also obtained. Based on the resultant values, if the probability of class 1 'Cl_1' is greater than the probability of class 2 'Cl_2' then the data distance d 'd' is classified to class 1 'Cl_1', else the data distance d 'd' is classified to class 2 'Cl_2'. In this manner, the data point gets allocated to the class with minimal variance. This in turn helps in performing efficient data classification for accurate decision making.

Experimental evaluation

The proposed algorithm was implemented using the MapReduce libraries in Hadoop. To assess the applicability of the proposed method, the method was applied to a benchmark dataset instacart market basket analysis dataset obtained from https://www.kaggle.com/c/instacart-market-basket-analysis/data. The instacart market basket analysis dataset is a relational set of files describing customers' orders over time. The goal of the competition was to predict which products will be in a user's next order. The dataset was anonymized and includes a sample of over 3 million grocery orders from more than 200,000 Instacart users. For each user, the dataset provided between 4 and 100 of their orders, with the sequence of products purchased in each order. The dataset also provides the week and hour of day the order was placed along with the relative measure of time between orders. Four entities, customer, product, order and aisle were included in the dataset with an associated unique id. Experimental evaluation were carried out on certain factors such as the feature selection rate, classification accuracy, classification time and error rate with respect to number of features and number of data points.

Impact of feature selection rate

The first and foremost metrics used in analysing the Big Data for predictive analysis is the feature selection rate. The rate at which the feature selection is made determines the efficiency of the method. Higher the feature selection rate, lower the efficiency is and lowers the feature selection rate, higher the efficiency is. In other words, feature selection rate refers to the time taken to select the features.

$$FSR = N * Time[FS] \tag{11}$$

From the above equation (11), the feature selection rate 'FSR' is measured based on the number of features considered for experimentation 'N' and the time consumed in feature selection '$Time[FS]$'. It is measured in terms of milliseconds (ms). Provided with 3 million grocery orders as sample (i.e. training dataset), experiments were conducted in the range of 5000 to 50000 features. The sample calculations for feature selection rate using the proposed TR-GIBMRC and existing evolutionary game theory-based method [1] and integrated artifacts [2] are given below.

Sample calculation for feature selection rate

Proposed TR-GIBMRC: With '5000' numbers of features considered for experimentation and the feature selection rate for single feature being '$0.013ms$', the overall feature selection rate is measured as given below.

$$FSR = 5000 * 0.013ms = 65ms$$

Existing Evolutionary game theory-based: With '5000' numbers of features considered for experimentation and the feature selection rate for single feature being '$0.018ms$',

the overall feature selection rate is measured as given below.

$$FSR = 5000 * 0.018ms = 90ms$$

Existing Integrated artifacts: With '5000' numbers of features considered for experimentation and the feature selection rate for single feature being '$0.022ms$', the overall feature selection rate is measured as given below.

$$FSR = 5000 * 0.022ms = 110ms$$

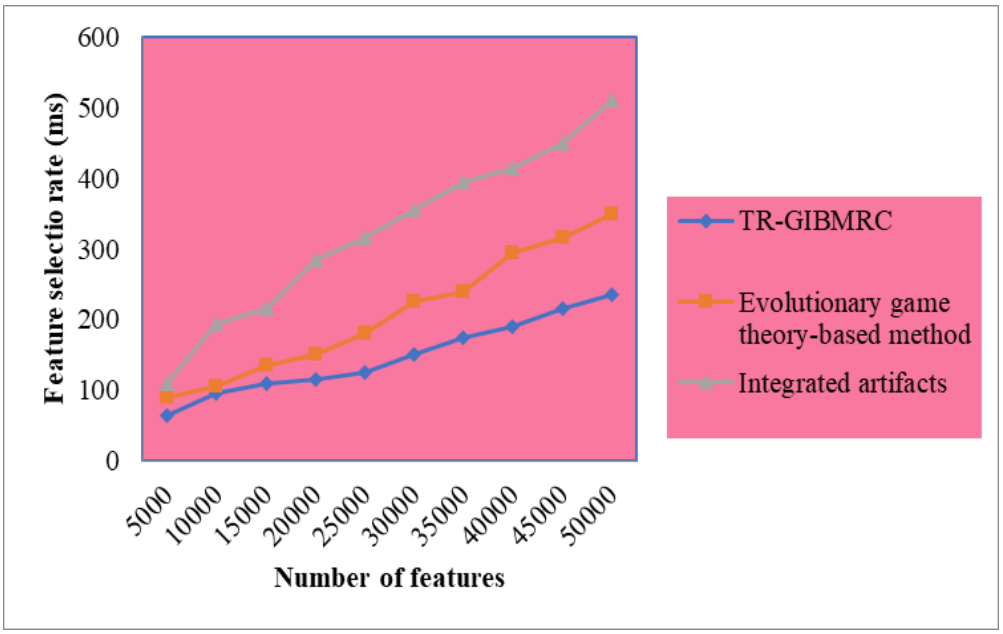

Figure 5 Performance results of feature selection rate using TR-GIBMRC, Evolutionary game theory-based method and Integrated artifacts

Figure 5 given above shows the convergence graph of feature selection rate. Here, x axis represents the numbers of features and y axis represents the feature selection rate. Different numbers of features in the range of 5000 to 50000 were used as sample set for conducting experiments collected from Instacart Market Basket Analysis. To measure the feature selection rate, with features being orders collected from orders.csv representing order_id, user_id, eval_set, order_number, order_dow, order_hour_of_day, days_since_prior_order are considered for experimentation. With the increase in the number of features, the feature selection rate also increases. This is because with higher the amount of features used for experimentation, the time consumed to select the feature also increases. However, comparative analysis shows better performance achieved in the TR-GIBMRC method. This is evidence from the sample calculations. With '5000' numbers of different features considered for experimentation at different time intervals, feature selection rate using TR-GIBMRC method was found to be '$65ms$', '$90ms$' using Evolutionary game theory-based [1] and '$110ms$' using Integrated artifacts [2] respectively. This is because of the application of Tobit Regressive Feature Selection model. By applying the Tobit Regressive Feature Selection model, highly associated or correlative features are selected. In other words, the Tobit Regressive Feature Selection being a statistical model infers the correlation between non-negative dependent variable and an independent variable for relevant feature selection. In this way, the feature selection rate using TR-GIBMRC method is reduced by 27% when compared to [1] and 53% when compared to [2] respectively.

Impact of classification accuracy

Classification accuracy is one of the most important metrics for measuring the predictive analysis. In other words, classification accuracy depends on the number of features correctly classified. It is mathematically formulated as given below.

$$CA = \frac{t}{N} * 100 \quad (12)$$

From the above equation (12), classification accuracy 'CA' refers to the percentage ratio of number of features (i.e. samples) correctly classified 't' to the overall features 'N' considered for experimentation. It is measured in terms of percentage (%). Provided with 3 million grocery orders as sample (i.e. training dataset), experiments were conducted in the range 5000 to 50000 features. The sample calculations for classification accuracy using the proposed TR-GIBMRC and existing evolutionary game theory-based method [1] and integrated artifacts [2] are given below.

Sample calculation for classification accuracy Proposed TR-GIBMRC: With ‘5000’ different numbers of features considered for experimentation, ‘4885’ numbers of features were correctly classified. Hence, the overall classification accuracy is measured as given below.

$$CA = \frac{4885}{5000} * 100 = 97.7\%$$

Existing Evolutionary game theory-based method: With ‘5000’ different numbers of features considered for experimentation, ‘4635’ numbers of features were correctly classified. Hence, the overall classification accuracy is measured as given below.

$$CA = \frac{4635}{5000} * 100 = 92.7\%$$

Existing Integrated artifacts: With ‘5000’ different numbers of features considered for experimentation, ‘4525’ numbers of features were correctly classified. Hence, the overall classification accuracy is measured as given below.

$$CA = \frac{4525}{5000} * 100 = 90.5\%$$

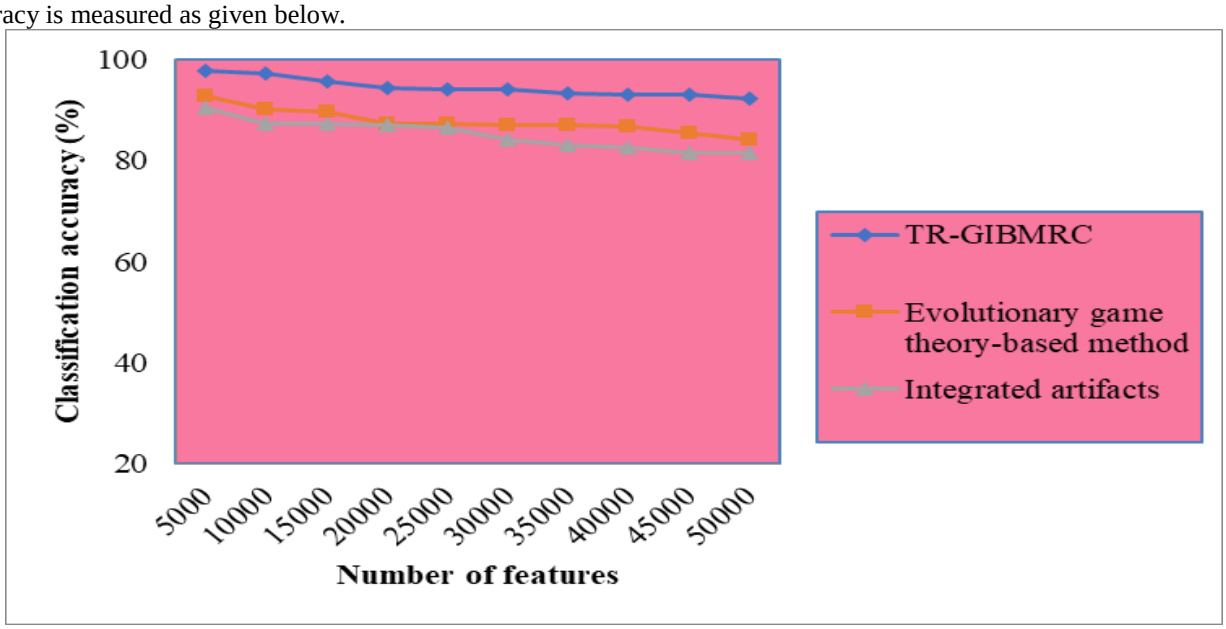

Figure 6 Performance results of classification accuracy using TR-GIBMRC, Evolutionary game theory-based method and Integrated artifacts

Figure 6 given above illustrates the graphical representation of classification accuracy (i.e. y axis) with x axis representing the different numbers of features in the range of 5000 to 50000 collected at different time intervals. Here, features considered for experimentation represents the sample submission made represented in sample_submission.csv denoting order_id and products. From the above sample calculations, with ‘5000’ different numbers of features (i.e. order_id) considered for experimentation, ‘4885’ numbers of features were correctly classified using TR-GIBMRC, ‘4635’ numbers of features were correctly classified using Evolutionary game theory-based method [1] and ‘4525’ numbers of features were correctly classified using Integrated artifacts [2]. With this, the classification accuracy using TR-GIBMRC was found to be ‘97.7%’, classification accuracy using Evolutionary game theory-based method [1] was found to be ‘92.7%’ and classification accuracy using Integrated artifacts [2] was found to be ‘90.5%’ respectively. From this it is evident that the classification accuracy was improved using the TR-GIBMRC method. This is because of the application of Gaussian Independence Bayes Map Reduce Classifier model. By applying the Gaussian Independence Bayes Map Reduce Classifier model, classification was made based on the strong independence assumptions between relevant features utilizing the Gaussian distribution. Due to this, the classification accuracy using TR-GIBMRC method was found to be improved by 8% compared to [1] and 11% compared to [2].

Impact of classification time

Classification time is yet another most important factor to be considered while obtaining the predictive analysis. In other words, classification time depends on the time taken to classify the features correctly. It is mathematically formulated as given below.

$$CT = N * Time\ [t] \quad (13)$$

From the above equation (13), classification time 'CT', refers to the product of the number of features considered for experimentation 'N' and the time consumed in correctly classifying the features '$Time\ [t]$'. It is measured in terms of milliseconds (ms). Provided with 3 million grocery orders as sample (i.e. training dataset), experiments were conducted in the range 5000 to 50000 features. The sample calculations for classification time using the proposed TR-GIBMRC and existing evolutionary game theory-based method [1] and integrated artifacts [2] are given below.

Sample calculation for classification time

Proposed TR-GIBMRC: With '5000' numbers of features considered for experimentation and the time consumed for classifying single feature being '$0.020ms$', the overall classification time is measured as given below.

$$CT = 5000 * 0.020ms = 100ms$$

Existing Evolutionary game theory-based method: With '5000' numbers of features considered for experimentation and the time consumed for classifying single feature being '$0.023ms$', the overall classification time is measured as given below.

$$CT = 5000 * 0.023ms = 115ms$$

Existing Integrated artifacts: With '5000' numbers of features considered for experimentation and the time consumed for classifying single feature being '$0.025ms$', the overall classification time is measured as given below.

$$CT = 5000 * 0.025ms = 125ms$$

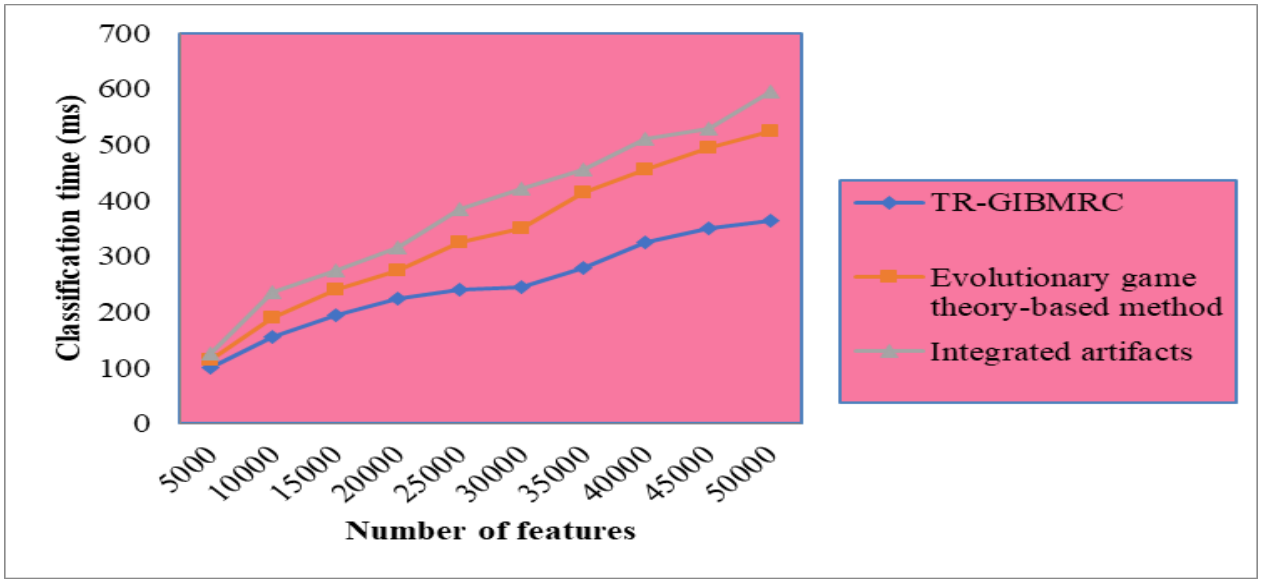

Figure 7 Performance results of classification time using TR-GIBMRC, Evolutionary game theory-based method and Integrated artifacts

Figure 7 given above shows the classification time for 50000 different numbers of features with features representing sample submission obtained from Instacart Market Basket Analysis dataset. From the figure, increasing the number of features, the classification time also increases. This is because of the reason that with the increase in the numbers of features, the same submission considered for experimentation also increases. Hence, the classification time for classifying different sample submission also increases. However, from the sample calculations it is evident that the classification time using the TR-GIBMRC method for '5000' numbers of features collected at different time intervals was found to be '$100ms$', classification time using Evolutionary game theory-based method was found to be '$115ms$' and the classification time using Integrated artifacts was found to be '$125ms$' respectively. The improvement in the classification time using the TR-GIBMRC method is due to the incorporation of Gaussian Independence Bayes Map classifier. By applying the Gaussian Independence Bayes Map classifier in the TR-GIBMRC method optimal selected features are segmented by measuring the mean and variance of data in each class. With this, the features to be classified are not only reduced by obtaining optimal relevant features, the time consumed in classifying the features is also found to be reduced. The improvement in classification time using TR-GIBMRC method was found to be 25% when compared to [1] and was found to be 34% when compared to [2].

Impact of error rate Finally, the classification error rate on an individual sample depends on the number of samples incorrectly classified to the overall samples considered for experimentation. It is measured as given below.

$$ER = \frac{f}{N} * 100 \quad (14)$$

From the above equation (14), the error rate 'ER' refers to the percentage ratio of incorrectly classified samples 'f' and the total samples considered 'N' for predictive analytics. The sample calculations for error rate using the proposed TR-GIBMRC and existing evolutionary game theory-based method [1] and integrated artifacts [2] are given below.

Sample calculations for error rate Proposed TR-GIBMRC: With '5000' different numbers of features considered for experimentation and '135' numbers of features incorrectly classified, the error rate is measured as given below.

$$ER = \frac{135}{5000} * 100 = 2.7\%$$

Existing evolutionary game theory-based method: With '5000' different numbers of features considered for experimentation and '195' numbers of features incorrectly classified, the error rate is measured as given below.

$$ER = \frac{195}{5000} * 100 = 3.9\%$$

Existing integrated artifacts: With '5000' different numbers of features considered for experimentation and '215' numbers of features incorrectly classified, the error rate is measured as given below.

$$ER = \frac{215}{5000} * 100 = 4.3\%$$

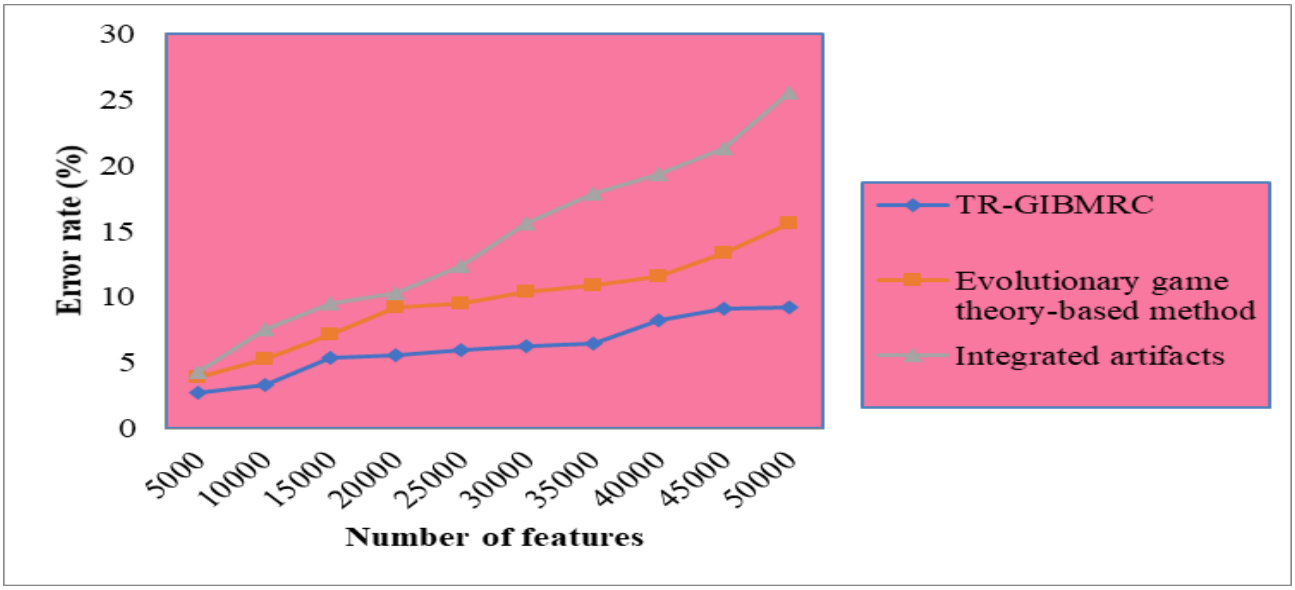

Figure 8 Performance results of error rate using TR-GIBMRC, Evolutionary game theory-based method and Integrated artifacts

Figure 8 given above shows the error rate (i.e. classification error rate) for 50000 different numbers of features considered. From the above figure it is evident that the error rate is found to be higher using the Integrated artifacts [2]. Comparative analysis however shows the error rate being in the reducing side using the TR-GIBMRC. With the application of the Gaussian Independence Bayes Map Reduce Classifier algorithm, prior of each class, likelihood of each class and posterior of each class and the mean and variance of each class were used to classify Big Data. With these five factors considered, the data point was found to be allocated to the class with minimal variance via aggregation of the intermediate key-value pairs. With this, the classification error rate was also found to be less using the TR-GIBMRC method. Besides, by applying Gaussian function, the error rate was found to be lessen using the TR-GIBMRC method by 35% compared to [1] and 54% when compared to [2].

IV. CONCLUSION

In this paper, Tobit Regressive based Gaussian Independence Bayes Map Reduce Classifier (TR-GIBMRC) method is proposed to classify the collected and stored data for each user's through efficient evolution of predictive analytics. The main goal of this proposed method is utilizing effective integrated feature selection and Map Reduce classification scheduling algorithm with the aiming to achieve maximum classification accuracy and minimum classification error rate on Data warehouse. For coarser construction on user requests, predictive analysis is improved by performing the Tobit Regressive based Gaussian Independence Bayes Map Reduce Classifier (TR-GIBMRC) method.

Then, the Tobit Regressive Log Likelihood algorithm is used select optimal relevant features with minimum classification time on the basis of identifying the relationship between the censored and uncensored data. The issue of classification of data points for predictive analysis addressed by introducing Gaussian Independence Bayes Map Reduce Classifier model which aids to enhance the feature selection and classification accuracy for features. The effectiveness of TR-GIBMRC method is estimated by attaining simulation results for testing the average feature selection rate, classification accuracy, classification time and error rate.

REFERENCES

1. Francesco Di Tria, Ezio Lefons and Filippo Tangorra, "A Framework for Evaluating Design Methodologies for Big Data Warehouses: Measurement of the Design Process", International Journal of Data Warehousing and Mining, Volume 14, Issue 1, January-March 2018, Pages 15-39 (Evaluating Design Methodologies for Big Data Warehouses)
2. Mohammad Karim Sohrabi and Hossein Azgomi, "Evolutionary game theory approach to materialized view selection in data warehouses", Knowledge-Based Systems, Elsevier, Volume 163, 2019, Pages 558–571 (Evolutionary Game Theory)
3. Nimmagadda Shastri L. Reiners Torsten and Wood Lincoln C, "On big data-guided upstream business research and its knowledge management", Journal of Business Research, Elsevier, Volume 89, 2018, Pages 143-158
4. Uthayasankar Sivarajah, Muhammad Mustafa Kamal, Zahir Irani, Vishanth Weerakkody, "Critical analysis of Big Data challenges and analytical methods", Journal of Business Research, Elsevier, Aug 2016
5. Venketesh Palanisamy, Ramkumar Thirunavukarasu, "Implications of big data analytics in developing healthcare frameworks –A review", Journal of King Saud University –Computer and Information Sciences, Dec 2017
6. Piotr Szul and Tomasz Bednarz, "Productivity frameworks in big data image processing computations - creating photographic mosaics with Hadoop and Scalding", ICCS 2014, 14th International Conference on Computational Science, Elsevier, Jun 2014
7. PekkaPääkkönen, DanielPakkala, "Reference Architecture and Classification of Technologies, Products and Services for Big Data Systems", Big Data Research, Elsevier, Feb 2015
8. Shadi Khalifaa, Patrick Martin, Rebecca Young, "Label-Aware Distributed Ensemble Learning:A Simplified Distributed Classifier Training Model for Big Data", Big Data Research, Elsevier, Nov 2018
9. Aurelle Tchagna Kouanou, Daniel Tchiotsop, Romanic Kengnea, Djoufack Tansaa Zephirin, Ngo Mouelas Adele Armelea, René Tchinda, "An optimal big data workflow for biomedical image analysis", Informatics in Medicine Unlocked, Elsevier, May 2018
10. Jing-Wei Liu, "Using Big Data Database to Construct New Gfuzzy Text Mining and Decision Algorithm for Targeting and Classifying Customers", Computers & Industrial Engineering, Elsevier, May 2018
11. Mehdi Sookhak, F. Richard Yu, Albert Y. Zomaya, "Auditing Big Data Storage in Cloud Computing Using Divide and Conquer Tables", IEEE Transactions on Parallel and Distributed Systems, VOL 29, NO 5, MAY 2018
12. Mohammadhossein Barkhordari, Mahdi Niamanesh, "Chabok: a Map-Reduce based method to solve data warehouse problems", Journal of Big Data, Springer, May 2018
13. Arantxa Duque Barrachina, Aisling O'Driscol, "A big data methodology for categorising technical support requests using Hadoop and Mahout", Journal of Big Data, May 2014
14. Yi-An Chen, Lokesh P. Tripathi, Kenji Mizuguchi, "TargetMine, an Integrated Data Warehouse for Candidate Gene Prioritisation and Target Discovery", Plos One, March 2011 | Volume 6 | Issue 3 | e17844
15. Marcello Ienca, Agata Ferretti, Samia Hurst, Milo Puhan, Christian Lovis, Effy Vayena, " Considerations for ethics review of big data health research: A scoping review "I", PLOS ONE | https://doi.org/10.1371/journal.pone.0204937 October 11, 2018
16. Ravi MadduriID, Kyle Chard, Mike D'ArcyID, Segun C. Jung, Alexis Rodriguez, Dinanath Sulakhe, Eric Deutsch, Cory Funk, Ben HeavnerID, Matthew Richards, Paul Shannon, Gustavo GlusmanID, Nathan Price, Carl Kesselman, Ian FosterI, "Reproducible big data science: A case study in continuous FAIRness", PLOS ONE | https://doi.org/10.1371/journal.pone.0213013 April 11, 2019
17. Lukman Ab. Rahim, Krishna Mohan KudiriID, Shiladitya Bahattacharjee, "Framework for parallelisation on big data", PLOS ONE | https://doi.org/10.1371/journal.pone.0214044 May 23, 2019
18. Hasna Njah1 Salma Jamoussi, Walid Mahdi, "Deep Bayesian network architecture for Big Data", Concurrency and Computation: Practice and Experience, Wiley Online Library, Nov 2017
19. JunPing Wang, WenSheng Zhang, YouKang Shi, ShiHui Duan, "Industrial Big Data Analytics: Challenges, Methodologies, and Applications", IEEE Transactions on Automation Science and Engineering, Jul 2018
20. Shan Suthaharan, "Big data analytics: Machine learning and Bayesian learning perspectives—What is done? What is not?", Data Mining and Knowledge Discovery, Wiley, May 2018

AUTHORS PROFILE

R.Sivakkolundu had pursued Bachelor of Computer Applications (BCA) from Madurai Kamarajar University in 2002, Madurai and Master of Computer Applications (MCA) from Bharathiar University, Coimbatore in 2005, Master of Philosophy in Computer Science from Bharathiar University, Coimbatore in 2011 and pursuing Ph.D in Computer Science from Bharathiar University, Coimbatore. Area of research is Data warehousing and Artificial Intelligence (AI). He published 2 papers in International Journals, presented 1 paper in National Conference.

Dr. V. Kavitha had pursued B.Sc., Computer Science from Bharathiar University in 1998, Coimbatore and Master of Computer Applications (MCA) from Bharathidasan University, Trichy in 2009, Master of Philosophy in Computer Science from Alagappa University, Coimbatore in 2005 and Ph.D in Computer Science from Karpagam University, Coimbatore in the year 2014. Area of research is Data Mining. Serving as a Reviewer and Editor in various International and National Journals. At present working as a professor in the Department of PG and Research Department of Computer Applications (MCA) at Hindusthan College of Arts and Science, Coimbatore-641 028. She published 37 papers in International Journals, presented 40 papers in International Conferences and National Conferences. She has 16 years of teaching experience and 10 yrs of Research experience.