Feature Selection using Stochastic Diffusion Search Algorithm in Big Data Analysis

Sumitra Srinivas K, Gangadhara Rao Kancharla

Abstract*: Big Data analysis has been viewed as the processing or mining of massive amounts of data used to retrieve information which is useful from large datasets. Among all the methods employed to deal with the analysis of Big Data, the selection of a feature is found extremely effective. A common approach which includes search making use of feature-based subsets which is relevant to the topic, tends to represent the dataset with its actual description. However, a search that makes use of such a subset is a combinatorial problem which is time-consuming. All commonly used meta-heuristic algorithms to facilitate feature choice. The Stochastic Diffusion Search (SDS) based algorithm has been a multi-agent global search algorithm based on agent interaction is simple to overcome combinatorial problems. The SDS will choose the feature subset for the task of classification. The Classification and Regression Tree (CART), the Naïve Bayes (NB), the Support Vector Machine (SVM) and the K-Nearest Neighbour (KNN) have been used to improve the performance. Results proved that the proposed method was able to achieve a better performance than existing techniques.

Index Terms*: Big Data Analysis, Feature Selection, Stochastic Diffusion Search (SDS) Algorithm, K-Nearest Neighbour (KNN) Classifier, Naïve Bayes (NB) Classifier, Classification and Regression Tree (CART) Classifier and Support Vector Machine (SVM) Classifier.

I. INTRODUCTION

Big Data refers to large datasets that may be complex or very large in such a way that technologies of data processing are found to be inadequate. There is a need for several technologies which may be used for storing and further processing the datasets that increase exponentially consisting of either structured, semi-structured or unstructured data. There are three characteristics that are defined in Big Data which are volume, variety and finally velocity. The volume is a large amount of data, variety signifies the data types and velocity is the data processing speed. The problems in the management of Big Data does not contain only the strings or numbers, but also includes social files, data from the web, video, audio, and geospatial data [1]. The primary issue faced by the technologies of Big Data was the provision of data processing of high quality along with data analysis. Recently, these technologies were applied in several fields of engineering and science which includes biomedical, biological and physical sciences. Further, these are also employed in large enterprises, financial corporations, and government agencies.

The Big Data can revolutionize all aspects of society, collection, and management of data which is useful since it is extremely challenging and complex. The expanding body contains hidden information found in a bulk of data which is non-traditional what has certain advanced technologies that are developed with an expert team which is multidisciplinary. The techniques of Machine learning having a power of computation in Big Data analytics. The focus of Machine learning on representing an input data along with the generalization patterns that are predicted in the future. There is a proper representation which may result in a performance that is high and in case the machine learning is found to be simple, it can result in a performance that is low. Thus, the primary element in machine learning is called feature engineering that is used for the purpose of construction of features and represents data from its raw input data. There are some major efforts needed in feature engineering and they are generally specific to domains. Machine learning is deployed widely for exploring the Big Data and its predictive features in several fields like search engines, Internet of Things (IoT) and medicine. In the direction of dealing with the analytics, there are some important machine learning sub-fields called deep learning that has been used for the extraction of useful data in Big Data [2].

The Hadoop has been found to be a software that is open source which was based on the architecture of Google MapReduce. This had been presented by Doug Cutting, an inventor of the Apache Lucene. This has been designed for processing large and complex data which was unstructured. It further provides a large amount of storage along with the ability of fast processing for tackling both tasks and jobs which are concurrent. At the time of having a massive set of data, it may become extremely challenging to process the data using a single computer. Thus it becomes important to divide massive data in the commodity computers. This type of parallel computing having some efficient along with a suitable framework has become the basic needs for Big Data processing. The Distributed File System (DFS) has managed data store in various computers falling in a cluster. As all these computers have been connected to one another by means of the network, the distributed file system has been identified to be very complex compared to its regular system. Thus, it becomes necessary to ensure rectification of problems that are connected to the network and manage the computer within the cluster [3].

Revised Manuscript Received on December 5, 2019.
***** Correspondence Author
Sumitra Srinivas K [PhD]*, Department of Computer Science, Sri Acharya Nagarjuna University, Guntur – 522510, Andhra Pradesh, India ksumisri@gmail.com.
Dr. Gangadhara Rao Kancharla, Department of Computer Science, Sri Acharya Nagarjuna University, Guntur – 522510, Andhra Pradesh, India, kancherla123@gmail.com.

The Map Reduce generally is implemented using a software framework that is open-source such as the Apache Hadoop used in distributed storage along with the processing of datasets of a large scale using computer clusters. The Apache Hadoop contains storage based on a Hadoop Distributed File System (HDFS) which includes a portion of processing such as Map Reduce. Thus, it may be implemented in various computations to tolerate the faults of hardware. Map functions then process input data for producing a new set of key-value pairs that are intermediate and composed through the master controller. An intermediate values tend to be assembled and are connected to the similar solutions agreed to similar machines [4]. Subsequently, the reduce function will work with the keys used one at a time and will associate the values related through the key which produce a set of values that are smaller which may result in their last key-value sets are the outputs. Outline of Map Reduce was employed in solving several problems in big data. For instance, the concept of Map Reduce was used to solve the problem of data reduction which was for filtering out all the unrepresented data was given from a dataset which was large.

Classification is a technique of data mining classifying data that is unstructured within a structured group or call helping the user in knowledge discovery. This classification ensures intelligent decision making with two different stages. The learning process and has datasets which are large, where analysis takes place after the execution where the accuracy is archived based on the patterns of classification. The two techniques found in classification are the supervised and the unsupervised. The former is also called the directed or the predictive classification and here the class is found in advance. The latter is called the undirected or the descriptive classification. This is a class that is unknown and once the classification is made a name may be assigned to the class [5].

There are large amounts of data which is high dimensional data which is pervasive in various domains which range from online education, transformation, health care, bioinformatics, e-commerce, and social media. At the time of data mining or a machine learning applications on the data that is high dimensional, the primary issue is called the curse of dimensionality. This indicates the data phenomenon which is sparser within a high dimensional space that may affect the algorithms adversely within a low dimensional space. Additionally, this type of existence of features that are of a high dimension there can be a significant demand on the requirements of memory storage of computation [6].

The selection of features taken as a technique of dimension reduction is quite effective in managing data of high dimension. It chooses relevant features directly in the construction of models. As the feature selection has a subset of the original features, the primary merit was that it was able to maintain all physical meanings in the original sets and provide better interpretability or readability [7]. Owing to this reason, it can be applied widely to various applications that are found in the real world like the analysis of gene and text mining. It obtains various relevant features by means of eliminating redundant and irrelevant features. Removing them brings down the cost of storage and computation with no major information loss or degradation of the performance of learning.

Generally, feature selection tends to employ a function of evaluation that identifies the feature subset. This evaluation function will measure the feature relevance for discriminating among target classes and this will depend on whether its evaluation function is able to interact with its classifier and the feature selection techniques was grouped into both filters and wrappers. The former will search and choose features on the basis of their power of discrimination without any interaction with its classifier. The methods make use of statistical scores to measure the power of discrimination which is effective. They include the Signal-to-Noise Ratio (SNR), the t-statistics (TS), and finally the F-test. Even though these filter methods are relatively simple and also efficient, they may not consider any correlation among features [8].

The latter, on the other hand, will encapsulate the method of search and the classifier in one single approach. It explores all feature space and subsets to choose an optimal subset which is able to achieve better performance and accuracy of classification compared to the other methods. There have been several investigations in the literature that have been made for solving the problem which includes the greedy search that was based on the Sequential Backward Selection (SBS) and the Sequential Forward Selection (SFS). The Evolutionary Computation (EC) have been considered well-known for their ability of global search employed for addressing the problem [9].

For the purpose of this work, the feature selection that makes use of the Big Data SDS algorithm has been employed. The rest of the investigation has been organized thus. The related work in literature has been explained in Section 2. The different methods used are talk over in Section 3. The experimental results are identified and discoursed in Section 4 and the conclusion was made in Section 5.

II. RELATED WORK

Peralta et al., [10] presented a new feature selection technique depends on an evolutionary computation using the paradigm of Map Reduce for obtaining various subsets. It decomposed its original dataset in instances that learn from its map phase and after this, the reduce phase had merged with the obtained results to form a final vector with feature weights permitting a flexible application of the process of feature selection making use of a threshold that determined the chosen subset features. This method of feature selection had been evaluated using three different classifiers in the structure of the Spark for addressing Big Data issues. In these experiments, there were 67 million instances and 2000 attributes thus proving this to be a framework that was well-suited in dealing with problems in Big Data. Zhao et al., [11] accessed a new context that had an efficient analysis of high dimensional feature selection that was original and distributed. This outline had a combination of variations in economic selection of features and the construction of econometric models that show the hidden patterns in economic development.

This functionality will rest on three different pillars which are: (i) the innovative techniques of data pre-processing for preparing economic data of high quality, (ii) a new and advanced result to distributed feature documents that locates the representative and important indicators from datasets that are multi-dimensional, and (iii) the new econometric models that are used for capturing all hidden patterns in economic development. The experiment based on economic data that was composed in Dalian in China had demonstrated that the framework has a superior performance in the examination of economic data.

Ramírez-Gallego et al., [12] proposed an extension of the mRMR to overcome the burden of computation. Being connected to a fast-mRMR, it has a package of three different implementations of the algorithm in various platforms such as the CPU for sequential execution, the Graphics Processing Units (GPU) used for parallel computing, and finally the Apache Spark used for the distributed computing that makes use of technologies of Big Data.

Yu et al., [13] had made a new presentation of the Scalable and Accurate OnLine Approach (SAOLA) in feature selection. There was a new theoretical analysis made on the bounds of correlations among features, and the SAOLA further employs a novel technique of pairwise comparison to maintain such a parsimonious model online. Also, in order to deal with all upcoming features, there was a new group-SAOLA algorithm that was proposed in the online group selection of features. This group-SAOLA algorithm was able to maintain online feature groups in sparse levels. There was another empirical work that was used based on real datasets showing two different algorithms, the SAOLA and the group-SAOLA.

Lin et al., [14] proposed an improved ICSO algorithm. The work was further applied to the ICSO algorithm for choosing all the select features in their experiment of text classification in the Big Data. The results proved that the ICSO performs better than the traditional CSO. In the classification of Big Data, results prove that making use of the Term Frequency-Inverse Document Frequency (TF-IDF) along with the ICSO for the purpose of feature selection is found to be much more accurate compared to making use of the TF-IDF alone.

Gu et al., [15] had made a new proposal that made use of a recently used variant of the PSO called the Competitive Swarm Optimizer (CSO) which was committed to an optimization of a large scale which solved problems in high dimensional selection of features. The CSO had been established for continuous optimization to accomplish feature selection which was taken to be a problem in combinatorial optimization. There was yet another archive technique that had been introduced and then compared to a canonical PSO along with a variant of the PSO and the proposed algorithm was able to choose smaller numbers in terms of features and they resulted in a better performance in terms of classification.

Sun et al., [16] had introduced another Lossless Pruned Naive Bayes (LPNB) classification algorithm which was personalized to the Big Data applications of the real world using many different classes. The LPNB was able to achieve a significant amount of speed ups by means of getting Information Retrieval (IR) based techniques. The LPNB was able to classify the text about eleven times faster compared to the Naïve Bayes in a real-world dataset that had about 7201 classes and higher gains that were extrapolated for taxonomies that were larger. In addition to this, it was lossless and the output was almost identical to the Naïve Bayes. This acceleration was not dependent on the structure of taxonomy and was used for flat and hierarchical taxonomies.

Hassanat [17] had made a proposal of another new approach based on feature vectors and their sorting in the training data found within a binary search tree which accelerated the classification of Big Data with the KNN approach. There were two methods employed for this. Both made use of the local points for sorting the examples that were based on the similarity they had to the local points. The local points were used for sorting examples on the basis of their comparison to global and extreme points, and the local points are considered in a random manner. The results of the experiment that were conducted on various Big datasets were able to show the rate of accuracy than the other state-of-the-art techniques and KNN classifier. This type of strong traits may be further used for improving the proposed method and its accuracy.

III. METHODOLOGY

A feature subset generation is a heuristic search technique which was the result of a choice of a candidate subset that is used for evaluation purposes. It employed strategies of search such as the complete, sequential and also the random search. An mRMR feature selection [18] was based on the filter that attempts at choosing the features which are relevant with the class labels and bring down any redundancy among the features that are chosen with a target class label and further brings down redundancy among the chosen features. It also makes use of Mutual Information (MI) to measure the similarity levels among two of the random and discrete variables. The Information Gain (IG) is dependent on information entropy [19] that has features providing this to the system. This becomes a parameter characterizing the substance found in thermodynamics to estimate the actual degree of chaos found within the system. In the year 1948, the father of information theory, C. E. Shannon, had applied the entropy to information and defined the actual amount of average information after this redundancy and had been excluded and called “information entropy”. This was a quantitative measure of the information for a certain degree of uncertainty for the random variables. For the purpose of this section, feature selection with SDS and the classifiers like SVM, CART, NB, and KNN were discussed.

A. Feature Selection Using Stochastic Diffusion Search (SDS) Algorithm

The SDS for the method of feature selection is making use of the direct communication of the agents that perform the evaluation of the feature subsets effectively. In the initial stages, every agent will be assigned to a new combination of a feature subset (or hypothesis) from its search space

(with various combinations of the features). The point to be noted here was that every agent makes use of the independent random split within the dataset in order to form two different subsets that are training and testing which have 80% and 20%. This hypothesis is a new binary string that will signifies the feature subset in the size of the subset. For this string, in case it is a bit which is 1, its equivalent feature is taken and 0 means it is not considered [20].

For the test phase, the activities of agents will be determined on the basis of its accuracy and will further be computed in its fitness function and here the agent will choose one more random agent and will compare his predictive accuracy accordingly in case the accuracy of the selecting agent that is higher than that of the random agent's and if not it is set to inactive. The process is continued for determining the status. The diffusion phase will commence.

During diffusion, the active, as well as the inactive agents will randomly pick yet alternative agent. When the inactive agent picks another agent, and the agent chosen randomly is active, it will offset the hypothesis and this is shared with an inactive agent or else, the choosing agent will pick up a random hypothesis from its search space subset. For the purpose of offsetting, there may be one feature that is removed randomly and one more is randomly added so conserving the size of the subset. Additionally, when there is an active agent picking another active agent, maintaining the same hypothesis, the choosing agent will be set to inactive and will be assigned a new and random hypothesis. This can free up the agents and improve diversity thus increasing the ability of the algorithm to search widely within the search space and this cycle of diffusion will be repetitive to as many numbers of iterations that are permitted.

The SDS algorithm used for the purpose of feature selection has been shown below:

1: *Initialisation phase*
2: *Assign agents to random hypotheses with inactive states*
3: *while less than iterations allowed do*
4: *#Evaluation Phase*
5: *for all agents do*
6: *Evaluate the fitness value*
7: *Find the* max *imum fitness value*
8: *end for*
9: *#Test Phase*
10: *for all agents do*
11: *if Agent's fitness > random agent's fitness then*
12: *Set agent as active*
13: *else*
14: *Set agent as inactive*
15: *end if*
16: *end for*
17: *#Diffusion Phase*
18: *for all agents do*
19: *if agent is inactive then*
20: *Select a random agent*
21: *if selected agent is active then*
22: *Copy its hypothesis & offset it*
23: *Calculate the fitness value*
24: *else*
25: *Pick a random hypothesis*
26: *Calculate the fitness value*
27: *end if*
28: *else #Agent is active*
29: *Select a random agent*
30: *if rand agent is active & shares hypothesis then*
31: *Set the selecting agent to inactive*
32: *Assign a new hypothesis*
33: *Calculate the fitness value*
34: *end if*
35: *end if*
36: *end for*
37: *end while*

B. K-Nearest Neighbour (KNN) Classifier

The KNN algorithm is an extremely simple one and is very easy to understand. The category of an unknown data point is identified with the Nearest Neighbour rule. The KNN will find the k nearest neighbour and this will be defined as a sample dataset. Since the KNN needs to store the entire training set, there may not be any redundancy within the training set for alleviating the problem [21].

For the KNN, a case will be grouped using the majority node and will be assigned to a class that is very common in the KNN which is measured using a distance function. In case the k is equal to one, the case will be allocated to the class belonging to its nearest neighbour. Several training trials kept within an n-dimensional space configuration along with an unknown sample in the KNN classifier which looks for a pattern space in the samples found closest to their unknown ones. This closeness has been defined based on the Euclidean distance or the distance between two different points, $X = (x_1, x_2, ..., x_n)$ and $Y = (y_1, y_2, ..., y_n)$ is (1):

$$d(X,Y) = \sqrt{\sum_{i=1}^{n} (X_i - Y_i)^2} \qquad (1)$$

The X and Y are the two objects and n denotes the actual number of their features. The KNN will work as:

- It finds the k parameter which is the nearest neighbour number.
- It calculates the distance between the training sample and query data.

- It then sorts the distance to find the nearest neighbour based on its kth minimal distance.
- It collects the category belonging to Y of its nearest distance.
- By employing the majority of the category belonging to the nearest neighbour, a prediction value for the instance of the query is found.

C. Naïve Bayes (NB) Classifier

An NB classifier has been simple in terms of the Statistical Bayesian Classifier. This is known as the Naïve and the assumption is that all the variables will contribute only to the mutually correlated classification. This is an assumption that is based on a class conditional independence. Independence Bayes, the Simple Bayes or the Idiot's Bayes are the other names. The probabilities of class membership are identified and this belongs to a certain class label. The NB classifiers will consider either the presence or the absence of a certain class has been distinct to the occurrence of the absence of any other class variable [22].

An NB classifier based technique that is on the basis of the Bayesian Theorem is used and this classification based on the theorem is as shown below: if X is the data sample that has a class label which is unknown, and H is a hypothesis, this data sample X may belong to a certain quantified class that is C. The Bayes theorem has been used for the purpose of computing its posterior probability which is (C|X), and from P (C), P (X), and P (X|C) as in equation (2).

$$P(C \mid X) = \frac{P(X \mid C).P(C)}{P(X)} \qquad (2)$$

Wherein,

P (C|X) denotes the posterior probability belonging to the target class.

P (C) is the prior probability of this class.

P (X|C) denotes the chances of being the predictor probability of a class.

P (X) denotes the class predictor's prior probability.

This NB classifier is functioned as:

- If D denotes a training set with class labels, every tuple stated using an n-dimensional element vector which is $X = (x1, x2, x3, ..., xn)$.
- Considering the number of classes are m that there are m classes C1, C2, C3...., Cm, it wants to group to an unknown tuple X, the classifier predicts that the X is part of the class that has a higher posterior probability and this is conditions on the X. Thus the Naïve Bayes will assign another new tuple X to a class in case P (Ci | X) > P (Cj | X) For the $1 \le j \le m$, and $i \ne j$, the above probabilities measured with the Bayes Theorem.

D. Classification and Regression Tree (CART) Classifier

The CART is yet another tree book learning method with a non-cluster conclusion yielding the grouping or the reversion trees. The CART makes use of the Gini-index for choosing the attribute as a measure of pollution. For splitting the records of the nodes, the feature that has the highest minimizing pollution has been employed. The CART has adapted data that has been numbered or with firm values to manage the feature ethics which were missed. The trimming of the complexity of cost has been used for generating the reversion trees [23].

After a forecast vector X is allocated, the CART will define a mutable distribution of Y. The supple process will apply a binary tree for splitting the space of prediction within the subsets where the Y distribution is even endlessly. The splitting rules will regulate the leaf nodules of the tree and is equivalent to all unlike separation based areas which link to every internal nodule. There is a trial of forecast that has been given with a single leaf nodule which moves from root to its leaf nodule for determining the Y distribution. If S is a specimen, the object feature which is $S_1, S_2, ..., S_k$ start from S, it will be based on classes as per equation (3):

$$Gini(S) = \sum_{i=1}^{K} \frac{|S_i|}{|S|}\left(1 - \frac{|S_i|}{|S|}\right) = \sum_{i \ne j} \frac{|S_i| \times |S_j|}{|S|^2} \qquad (3)$$

E. Support Vector Machine (SVM) Classifier

The SVM was a supervised learning process used in regression and classification issues that depend on the actual output. The SVM makes use of a statistical learning theory that maximizes the generalization property of a classifier model which is generated. The SVM will prevent overfitting [24]. The theory of Statistical learning has a value of fitting of its training data. The observed risk was $R = \frac{1}{n}\sum_{i=1}^{n} l(f_\theta(x_i), y_i)$, and this is the average loss of 1 for the selected estimator above its training set $(x_i y_i)$. The SVM further makes use of a new set of training data which calculates every given input either of the two potential classes {-1; 1}. Here a hyperplane will be defined by the $w^T x + b = 0$, in which, $w \in R^n$ denotes an orthogonal of the hyperplane and $b \in R^n$ will be the bias. With training data D, there is a point as given below (4):

$$D = \{(x_i, y_i) \mid x_i \in R^m,\ y_i \in \{-1, 1\}\}_{i=1}^{n} \qquad (4)$$

Here x_i denotes a real vector that is m-dimensional and y_i is the input vector of x_i, either -1 or 1. The purpose of SVM is searching a new hyperplane which increases the boundary existing in the middle of both class samples of the D having an empirical risk which is small. For the new generalization property of the SVM, there are two of the hyperplanes that have been defined so that $w^T x + b = 1$ and the $w^T x + b = -1$. Both functions may be simplified to form (5):

$$y_i(w^T x_i + b) \ge 1 \qquad (5)$$

An SVM will aim at maximizing the actual distance between both the hyperplanes. The distance between them will be $\frac{1}{\| w \|}$. Training of an SVM for a case that is non-separable has been solved by making use of the problem of quadratic optimization as per equation (6).

$$\min imize: P(w,b,\xi) = \frac{1}{2}\| w \|^2 + C\sum_{i=1}^{m}\xi_i$$
$$subject\ to: y_i((< w,\phi(x_i) > +b)1-\xi_i \quad (6)$$
$$\xi_i \geq 0$$

For $i = 1,....,m$, wherein the ξ_i are the slack variables and the C denotes the cost variable for every slack, C denotes the control factor for the maximization of margin and the minimization of empirical risk. The SVM and its decision function will be f(x) =w^T ϕ(x) + b, wherein w and b will be computed based on the problem of optimization P as per equation (6). By making use of the Lagrange multipliers, P the problem optimization in equation (6) is shown in (7):

$$\min: F(\alpha) = \frac{1}{2}\alpha^T Q\alpha^T - \alpha^T 1$$
$$Subject\ to: 0 \leq \alpha \leq C \quad (7)$$
$$y^T\alpha = 0$$

Wherein $[Q]_{ij} = y_i y_j \phi^T(x_i)\phi(x_j)$ denotes its Lagrangian multiplier variable. The kernel function is represented as a kernel function as $K(x_i, x_j) = \phi^T(x_i)\phi(x_j)$. So $[Q]_{ij} = y_i y_j K(x_i, x_j)$.

A hyper-plane will divide the dataset into two. If they are separable non-linearly, the kernel functions are employed. For the general division of the kernel functions, two kernel functions called local and global are considered. The former has the function data points that are adjacent to one another for making an impact on the kernel points. The latter will be in distance from one another and can make a major influence on the kernel points [25].

All the kernel functions are:

The Linear Kernel Function: $K(x_i, x_j) = 1 + x_i^T x_j$

The Polynomial Kernel Function: $K(x_i, x_j) = (1 + x_i^T x_j)^p$

The Radial Basis Kernel Function: $K(x_i, x_j) = \exp(-\gamma \| x_i - x_j \|^2)$

The Exponential Radial Basis Kernel Function: $K(x_i, x_j) = \exp\left(-\frac{\| x_i - x_j \|}{2\sigma^2}\right)$

The Gaussian Radial Basis Kernel Function: $K(x_i, x_j) = \exp\left(-\frac{\| x_i - x_j \|^2}{2\sigma^2}\right)$

The Sigmoid Kernel Function: $K(x_i, x_j) = \tanh(kx_i^T x_j - \delta)$

The Radial Basis Function (RBF) has better results on being compared to the linear or the polynomial kernel function. Main difficulty is to select the optimum value of a penalty factor C and the kernel factor which σ with a better accuracy.

IV. RESULTS AND DISCUSSION

Cifar-10, 5 classes are considered using 4000 images per class (airplane, bird, ship, truck and automobile). The SDS-KNN, SDS–NB, SDS-CART and SDS-SVM techniques are considered. Investigations are assessed via class A to class E signifying airplane, bird, ship, truck and automobile respectively. The true positive rate, positive predictive value, f measure and misclassification rate are represented in figures 1 to 4:

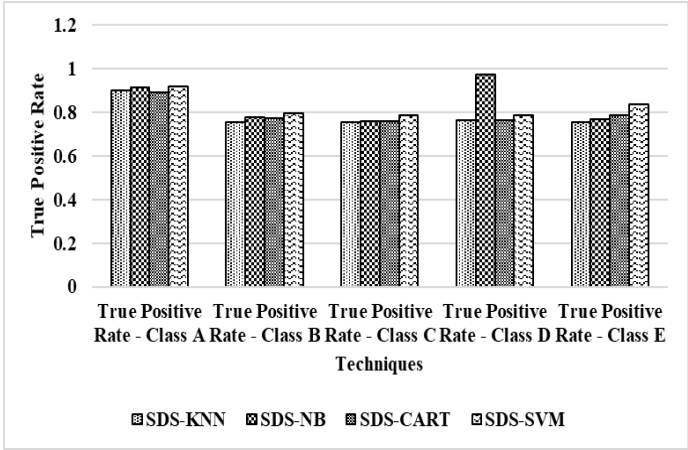

Figure 1 True Positive Rate for SDS-NB

Figure 1 shows that the SDS-NB takes higher average true positive rate by 6.46%, 5.42% and 1.56% for SDS-KNN, SDS-CART and SDS-SVM respectively.

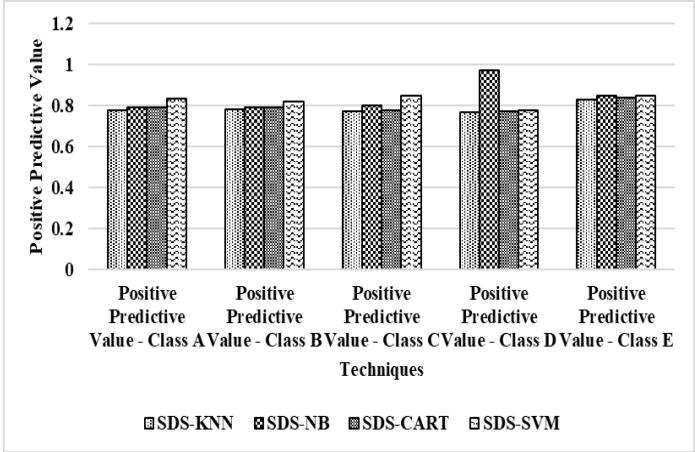

Figure 2 Positive Predictive Value for SDS-NB

Figure 2 shows that the SDS-NB takes higher average positive predictive value by 6.56%, 5.58% and 1.74% for SDS-KNN, SDS-CART and SDS-SVM respectively.

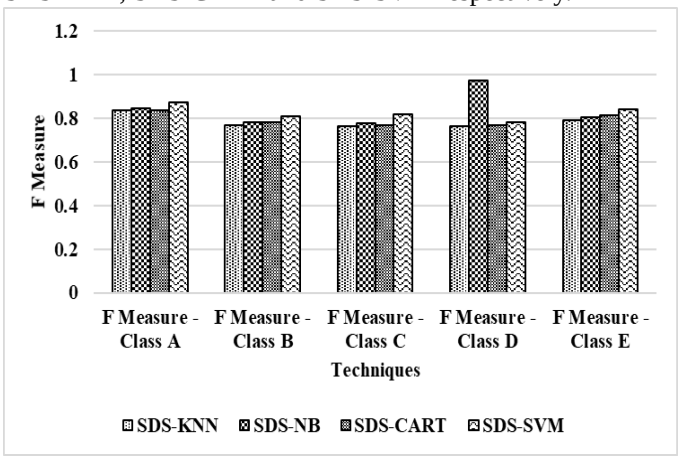

Figure 3 F Measure for SDS-NB

Figure 3 shows that the SDS-NB takes higher average f measure by 6.52%, 5.44% and 1.57% for SDS-KNN, SDS-CART and SDS-SVM respectively.

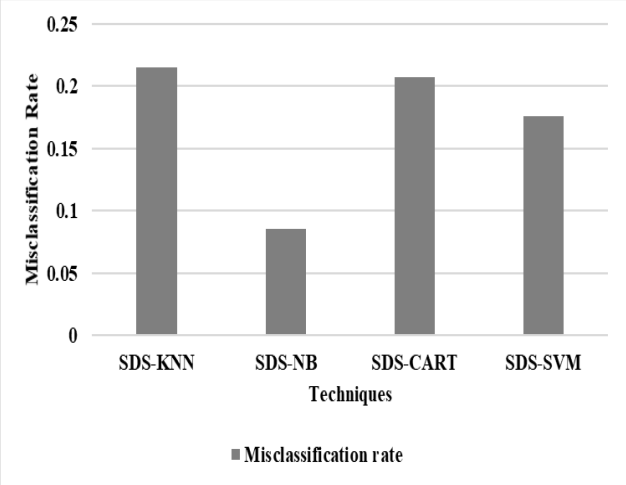

Figure 4 Misclassification Rate for SDS-NB

Figure 4 shows that the SDS-NB takes lower misclassification rate by 86.38%, 83.19% and 69.22% for SDS-KNN, SDS-CART and SDS-SVM respectively.

V. CONCLUSION

The Big Data has some massive data with data heterogeneity and data streams. The database dimension had been reduced using techniques of feature selection. The method of novel selection having an SDS algorithm had been proposed. The SDS had provided a balance in the middle of its local and global examination by repeating over the phases of test and diffusion. It also combines the offsetting along with a recruitment strategy that was context sensitive. The KNN classifier is a lazy learning algorithm which was efficient and the NB classifier needs a short time of computation and to enhance the classification presentation via removing all irrelevant features. The CART algorithm identifies all significant variables to eliminate the ones that are non-significant. The SVM will balance both accurate and proper data to find a compromise between the complexity and error of the classifier. The results have shown that an SDS-NB could get a higher rate of true positive by about 6.46% for the SDS-KNN, by about 5.42% for the SDS-CART and by about 1.56% for the SDS-SVM respectively. In the same way, the SDS-NB has an average positive predictive value that was higher and an f measure by the SDS-KNN, the SDS-CART and finally the SDS-SVM respectively. However, the SDS-NB has a misclassification rate that was lower by the SDS-KNN, the SDS-CART, and the SDS-SVM respectively.

REFERENCES

1. Demidova, L., Nikulchev, E., & Sokolova, Y. (2016). Big data classification using the SVM classifiers with the modified particle swarm optimization and the SVM ensembles. International Journal of Advanced Computer Science and Applications, 7(5), 294-312.
2. Jan, B., Farman, H., Khan, M., Imran, M., Islam, I. U., Ahmad, A., ... & Jeon, G. (2017). Deep learning in big data Analytics: A comparative study. Computers & Electrical Engineering. 1-13.
3. Adhikari, B. K., Zuo, W. L., Maharjan, R., & Guo, L. (2018, November). Sensitive Data Detection Using NN and KNN from Big Data. In International Conference on Algorithms and Architectures for Parallel Processing (pp. 628-642). Springer, Cham.
4. Tsai, C. F., Lin, W. C., & Ke, S. W. (2016). Big data mining with parallel computing: A comparison of distributed and MapReduce methodologies. Journal of Systems and Software, 122, 83-92.
5. Koturwar, P., Girase, S., & Mukhopadhyay, D. (2015). A survey of classification techniques in the area of big data. arXiv preprint arXiv:1503.07477.
6. Li, J., & Liu, H. (2017). Challenges of feature selection for big data analytics. IEEE Intelligent Systems, 32(2), 9-15.
7. Tsamardinos, I., Borboudakis, G., Katsogridakis, P., Pratikakis, P., & Christophides, V. (2018). A greedy feature selection algorithm for Big Data of high dimensionality. Machine Learning, 1-54.
8. Wang, Y., Ke, W., & Tao, X. (2016). A feature selection method for large-scale network traffic classification based on spark. Information, 7(1), 6.
9. Alanazi, N. N. S. (2017). An Intelligent Hybrid Feature Selection for Big Data Analytics: Application to Microarray Data Analysis. International Journal of Engineering Research & Technology (IJERT), 7 (5), 89-95.
10. Peralta, D., del Río, S., Ramírez-Gallego, S., Triguero, I., Benitez, J. M., & Herrera, F. (2015). Research Article Evolutionary Feature Selection for Big Data Classification: A MapReduce Approach. 1-11.
11. Zhao, L., Chen, Z., Hu, Y., Min, G., & Jiang, Z. (2018). Distributed feature selection for efficient economic big data analysis. IEEE Transactions on Big Data, 4(2), 164-176.
12. Ramírez-Gallego, S., Lastra, I., Martínez-Rego, D., Bolón-Canedo, V., Benítez, J. M., Herrera, F., & Alonso-Betanzos, A. (2017). Fast-mRMR: Fast Minimum Redundancy Maximum Relevance Algorithm for High-Dimensional Big Data. International Journal of Intelligent Systems, 32(2), 134-152.
13. Yu, K., Wu, X., Ding, W., & Pei, J. (2016). Scalable and accurate online feature selection for big data. ACM Transactions on Knowledge Discovery from Data (TKDD), 11(2), 16.
14. Lin, K. C., Zhang, K. Y., Huang, Y. H., Hung, J. C., & Yen, N. (2016). Feature selection based on an improved cat swarm optimization algorithm for big data classification. The Journal of Supercomputing, 72(8), 3210-3221.
15. Gu, S., Cheng, R., & Jin, Y. (2018). Feature selection for high-dimensional classification using a competitive swarm optimizer. Soft Computing, 22(3), 811-822.
16. Sun, N., Sun, B., Lin, J. D., & Wu, M. Y. C. (2018). Lossless Pruned Naive Bayes for Big Data Classifications. Big Data Research, 14, 27-36.
17. Hassanat, A. B. (2018). Two-point-based binary search trees for accelerating big data classification using KNN. PloS one, 13(11), e0207772.
18. Gulgezen, G., Cataltepe, Z., & Yu, L. (2009). Stable and Accurate Feature Selection. Springer
19. Ding, J., & Fu, L. (2018). A Hybrid Feature Selection Algorithm Based on Information Gain and Sequential Forward Floating Search①. Journal of Intelligent Computing Volume, 9(3), 93.
20. Alhakbani, H., & al-Rifaie, M. M. (2017, July). Feature selection using stochastic diffusion search. In Proceedings of the Genetic and Evolutionary Computation Conference (pp. 385-392). ACM.
21. Thakor, H. R. (2017). A Survey Paper on Classification Algorithms in Big Data. International Journal Of Research Culture Society, 1 (3), 21-27.
22. Jadhav, S. D., & Channe, H. P. (2016). Comparative study of K-NN, naive Bayes and decision tree classification techniques. International Journal of Science and Research, 5(1), 1842-1845.
23. Manikandan, R. P. S., & Kalpana, A. M. (2017). Feature selection using fish swarm optimization in big data. Cluster Computing, 1-13.
24. Çatak, F. Ö., & Balaban, M. E. (2016). A MapReduce-based distributed SVM algorithm for binary classification. Turkish Journal of Electrical Engineering & Computer Sciences, 24(3), 863-873.
25. Priyadarshini, A. (2015). A map reduce based support vector machine for big data classification. International Journal of Database Theory and Application, 8(5), 77-98..