Available online at: http://www.ijmtst.com/vol4issue11.html

International Journal for Modern Trends in Science and Technology

ISSN: 2455-3778 :: Volume: 04, Issue No: 11, November 2018

Prediction Classifier using Support Vector Machine for Big Data Computation in Cloud Environment

Bhallamudi Ravikrishna | Dr. Harsh Pratap Singh

Research Scholar, Computer Science and Engineering, Sri Satya Sai University of Technology and Medical Sciences, Sehore (M.P)
ravikrishnabh@gmail.com

Associate Professor, Department of Computer Science and Engineering, Sri Satya Sai University of Technology and Medical Sciences, Sehore (M.P)
harshaprathapsingh01@gmail.com

To Cite this Article
Bhallamudi Ravikrishna and Dr. Harsh Pratap Singh, "Prediction Classifier using Support Vector Machine for Big Data Computation in Cloud Environment", *International Journal for Modern Trends in Science and Technology*, Vol. 04, Issue 11, November 2018, pp.-21-27.

Article Info
Received on 21-Oct-2018, Revised on 19-Nov-2018, Accepted on 26-Nov-2018.

ABSTRACT

The major challenges in big data applications are the availability of enormous volumes of data and extraction of expensive information or knowledge for future actions. The distributed data mining on cloud data consists of minimum computational overhead and communication costs. The distributed environments are enough for utilizing the larger datasets. With the presence of large datasets, the conventional classification approaches have failed to produce the desirable results. The classification performance is not effective for distributing and sharing the information along with different applications. DSV-CP model is proposed to offer efficient computation on big data applications and allocation of information in cloud computing environment. Initially, preprocessing is performed in DSV-CP model based on IED that helps to remove the noise and inconsistent data that are obtained from various sources. While removing the noise and inconsistency present in the data, computation time and space complexity involved during information sharing in cloud environment are reduced. DSV-CP model uses support vector prediction classifier to classify the data based on the users' query request using parallel hyperplanes to improve the classification accuracy of the user request information on big data. Finally, DSV-CP model accurately identifies the user request information on big data with the help of the classified data.

I. INTRODUCTION

Big data applications share sensitive information with multiple servers by protecting the data known as data privacy. Bank transactions, medical records for a particular patient and other data sharing are referred to as sensitive information in information sharing [1]. Data privacy protection is considered by the following two approaches. Initially, limiting data entries are made into the server by adding some other documents in information or managing the data given by the users. Then the sensitive information is located in the records from anonymized data. Domain and application of knowledge identify necessary information using big data analytical technique. Based on different domain applications, data

privacy and information sharing between various servers are considerably different. Different mining algorithms are introduced for sharing the information between servers in computing network. Autonomous sources and decentralized control are the characteristics of big data with mining algorithm to aggregate the disseminated data for providing transmission privacy[2]. They collect different data from distributed sites by carrying mining behavior with biased model in computing network. The different features of big data applications with mining algorithms are spare, uncertain and incomplete data. In this, spare data reduces the complexity and consistency on sharing the information. Structured data, unstructured data and semi-structured data are the different types of data used for sharing complex heterogeneous data [3]. Multiple information sources ensure pattern matching with mining algorithm to process the information with three different levels, namely data level, model level and knowledge level. Basic data sources achieve data distribution at data level with a global view by estimating statistical variation between different user knowledge data. Data are localized with local mining algorithm based on model or pattern level to interchange the data with multiple users. Finally, the knowledge level process is carried out for determining the correlated data from autonomous sources that are generated from various users, but the unstructured data are not linked with complex heterogeneous data for sharing different information between the users.

2. FLEXIBLE ANALYTIC PLACEMENTS WITH QUANTITATIVE MODEL:

Data analytics are placed with I/O path for reducing the data movement by introducing Peta scale computation. Simulation, analytics and visualization are included in specific applications for executing the computing nodes[4]. After the execution, I/O node is used as the storage device. Similarly, large data movements are produced between computing node and staging node when data analytics is located on staging nodes. Data queries are requested from the user nodes to show the performance of different data analytics by developing analytics algorithm during flexible framework. Quantitative model uses different requests from computing nodes on data analytics[5] to reduce the performance cost and shows reduction of data. Different methods are provided for transferring the data to destination through I/O path, and they are mainly classified as memory-to-memory transfer and memory-to-storage transfer [6]. However, when bandwidth for data transfer is higher, optimized data compression is not effective. Therefore, support vector prediction classifier is designed in DSV-CP model to classify the data based on user query. Classified queries accurately identify the user information on big data with the help of the classified data. DSV-CP MODEL [7] Big data applications are used for sharing the structured and unstructured information by collecting the data effectively to achieve faster response and reduced classification time. The design of DSV-CP model is to perform big data computation and information sharing in cloud computing environment. In this, big data represents enormous amount of data in different formats that are handled by the established databases. Improvements in classification accuracy and prediction accuracy of user request information on big data are the main purposes for developing the proposed DSV-CP model based on the present data and historical facts in the cloud. Big data application collects the information to capture, process and manages the sharing or distribution of the information among various resources[8]. The architecture of the proposed DSV-CP model is given in Figure 1.

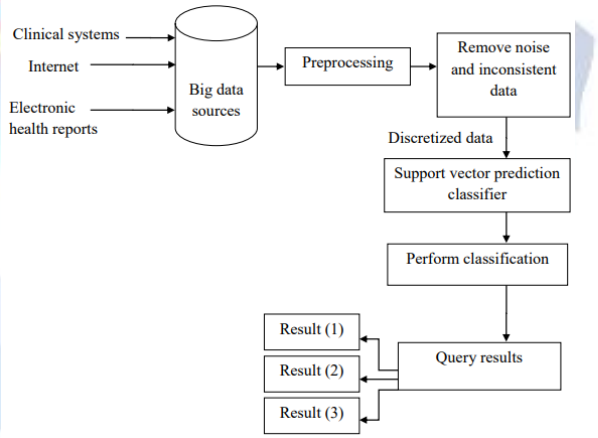

Figure 1 Architecture of DSV-CP Model

The proposed DSV-CP model performs computing and sharing of information on big data. The initial step in the design of DSV-CP model is the data preprocessing where the data are discretized aiming at reducing the computation time and space complexity [9]. In the next step, it employs the support vector prediction classifier with the discretized data for its corresponding query results obtained from the user requests. As shown in Figure 1, the DSV-CP model initially obtains the input from different sources, EHR, clinical

systems, internet and so on in various formats, i.e., flat files, .CSV, ASCII, tables and so on, from various locations.

3. SUPPORT VECTOR PREDICTION CLASSIFIER

In the proposed DSV-CP model, support vector prediction classifier helps to identify different sets of big data each classified with various types of data. Let an example of big data from different sources, namely clinical system [10], internet, EHR and from other mobile devices be considered. Figure 2 shows the support vector prediction classifier employed in the proposed DSV-CP model for efficient classification of big data.

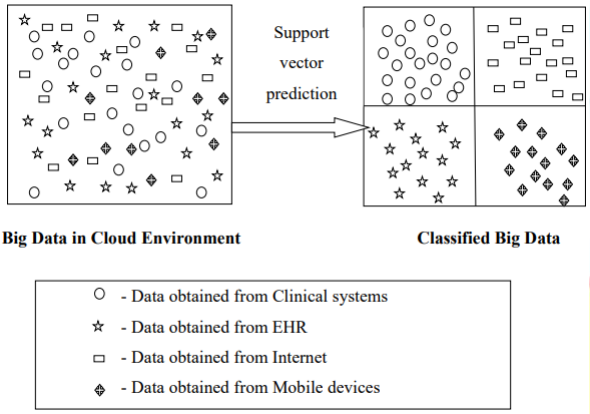

Figure 2 Classification of Big Data using Support Vector Prediction Classifier

Let a bank marketing data which includes personal details, contact and other information be considered. The bank marketing data contains both numeric and text attributes. The transaction process is carried out by the bank account given by the person and the data presented in the account is considered with variable attributes. Data transaction is indicated with its frequency for improving the prediction accuracy [11], but to classify and to identify the solution to data, rule mining algorithm is required. Therefore, in the proposed DSV-CP model, the support vector prediction classifier is used. Support vector prediction classifier Classified Big Data – Data obtained from Clinical systems – Data obtained from I– Data obtained from Internet – Data obtained from Mobile devices Big Data in Cloud Environment This classifier reduces the misclassification errors and improves the search accuracy of user request information [12], i.e., classification accuracy of big data.

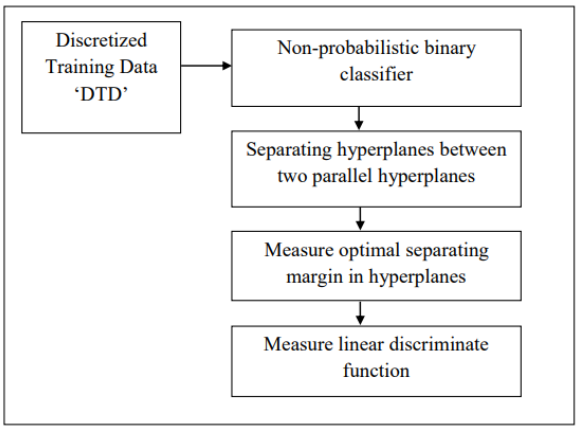

Figure 3 Block Diagram of Support Vector Prediction Classifier

Support vector prediction classifier used in the proposed DSV-CP model is shown in Figure 3. It takes the preprocessed discretized data as input and predicts the data for each user query request. It separates two possible classes from the input, making the support vector prediction classifier a non-probabilistic binary linear classifier [13]. In the proposed DSV-CP model, the support vector non-probabilistic binary classifier constructs a set of hyperplanes that are split into two types of representations such as linear and non-linear support vector. To start with, as shown in Figure 3, the proposed DSV-CP model considers non-probabilistic binary classifier with an input of Discretized Training Data (DTD) [14] as given in Equation (1).

$$DTD = (p_1, q_1), (p_2, q_2), \ldots, (p_n, q_n) \quad (1)$$

In the Equation (1), 'p and q' are two separate possible classes. The class '$p_i \in \{-1, 1\}$' denotes the class to which 'p_i' belongs to the dimension of a space and the number of users in the cloud environment. Similarly, 'q' is denoted as a scalar vector. Two parallel hyperplanes are created in cloud environment on each side of the hyperplanes which divide the data. The separating hyperplanes are the hyperplanes that exploit the interval between the two parallel hyperplanes [15]as shown in Figure 4.

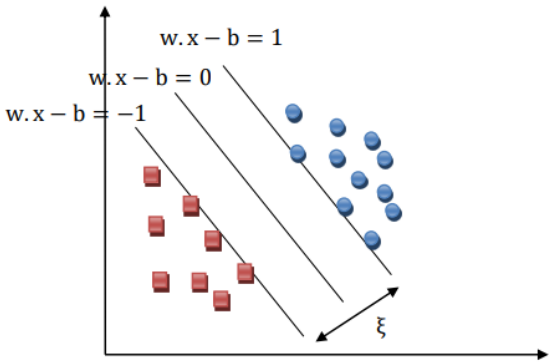

Figure 4 Hyperplanes for a Support Vector Trained with Samples from Two Classes

Figure 4 shows the maximum margin hyperplanes for a support vector prediction classifier trained with samples from two classes '(pi ,qj)'. When the training data are linearly separable, two parallel hyperplanes are selected for separating the two classes of data. In the Figure 4, 'w' is the normal vector to the hyperplane and 'b' determines hyperplane from origin along the normal vector. In this, 'x' indicates the classes and each class is referred to as p-dimensional data.The optimal separating margin in hyperplanes for a support vector prediction classifier is identified in the proposed DSV-CP model that obtains the slack variables 'ξ'. When the slack variable is greater than one, the point is said to be a misclassified point[16][17]. The proposed hyperplane does not exceed the point. Therefore, the slack variables in the proposed DSV-CP model are introduced in such a way that the misclassification errors are reduced.

$$Min, \left\{ \frac{1}{2}|w|^2 + C \sum_{i=1}^{n} \xi_i \right\} \qquad (2)$$
$$subject\ to\ q_j(w^n p_i + f)$$

In Equation (2), the slack variables represented as 'ξi ' correspond to 'w' indicating the vector and 'f' standing for a scalar value. The classification constant is denoted as 'C' in the Equation (2). With the obtained optimization of cloud data, the proposed DSV-CP model improves the classification accuracy[18][19]. In this, a scaling is essential to safeguard against variable, i.e., attributes, with larger difference. The proposed model examines the training data by separating hyperplanes using the mathematical Equation (3).

$$p.x + q = 0 \qquad (3)$$

Equation (3) represents the separation on hyperplanes where 'q' is denoted as scalar and 'p' is denoted p-dimensional vector and the vector 'p' is perpendicular to the separating hyperplanes. Discretization factor 'x' is located for removing the noise and inconsistent data to provide better information sharing of big data. When the offset parameter 'q' is inserted in hyperplanes, it permits the optimization of the separating margin[20][21]. Therefore, parallel hyperplanes, with the objective of obtaining the big data from various sources and data sharing in a parallel manner, are described in Equations (4) and (5).

$$p.x + q = 1 \qquad (4)$$
$$p.x + q = -1 \qquad (5)$$

If the training data in the proposed DSV-CP model are linearly separable, parallel hyperplanes are chosen so that there are no points connecting them, and subsequently, their distance is maximized, i.e., to improve classification accuracy and prediction rate. In contrast, to reduce the misclassification errors, Lagrangian multiplier 'β_j' is obtained according to the Karush-Kuhn-Tucker condition 'α_i '. If 'α_i > 0', then the corresponding data 'p_i' and the proportion of training data 'q_j' are called the support vector[22][23]. The function of linear combination with optimal separating margin is given by 'f '. Therefore, the Linear Discriminate Function (LDF) with optimal separating margin is expressed as in Equation (6).

$$LDF = \left(\sum_{i=1}^{n} \alpha_i q_j p_i + f \right) \qquad (6)$$

Algorithm 1 Process of Support Vector Prediction classifier

Input: Discretized Training Data '
$DTD = (p_1, q_1), (p_2, q_2), \ldots, (p_n, q_n)$'
Output: Classification of Big Data
Step 1: **Begin**
Step 2: **For each** Discretized Training Data 'DTD'
Step 3: Measure optimal separating margin in hyperplanes
Step 4: Measure linear discriminate function with optimal separating margin
Step 5: **End for**
Step 6: **End**

As shown in Algorithm 1, the support vector prediction classifier for the discretized data initially separates the hyperplanes for examining the training data. Then the classifier describes parallel hyperplanes to restrict hyperplanes from passing through the origin. It verifies if the training data are

linearly separable to maximize their interval, and subsequently, it determines the interval between the hyperplanes to identify the right hyperplanes for classifying the data. Finally, this classifier selects optimal hyperplanes to classify the data which results in improved search accuracy on big data, i.e., classification accuracy[24]. Thus, DSV-CP model improves the prediction accuracy in a significant manner.

4. EXPERIMENTAL EVALUATION PERFORMANCE ANALYSIS OF DSV-CP MODEL

The DSV-CP model with efficient big data computation and information sharing among cloud instances is implemented, which provides resizable computing capacity in cloud. It uses the HDFS two-layer namespace to minimize the computational complexity and computation cost. Cloud computing services first recognize the users' requests for information sharing, then taking best decisions from the data, they pass the information onto the other users without redundancy. The information sharing with big data is done in an efficient manner in cloud computing environment with HDFS two-layer namespace. The performance of DSV-CP is evaluated using Stanford Large Network dataset collection that uses Amazon product co-purchasing network for conducting experiments in an extensive manner. Certain attributes used in Amazon product co-purchasing network are listed in Table 1.

Attributes	Description
N	Number of nodes in social network
n_s	Number of senders of recommendations
n_r	Number of recipients of recommendations
r	Number of recommendations
e	Number of edges in social network
V	Number of reviews of the product
T	Average product rating

Table 1 Product Attributes with Description

The performance evaluation of the proposed DSV-CP and existing methods, namely DM-BD and Flex-Analytics are conducted with Amazon EC2 dataset collections, to provide effectiveness to big data computation. It is compared with existing methods DM-BD and Flex-Analytics. Cloud-Sim simulator is also used to measure the experiment parameters and the results are presented for different sizes of big data based on the user requests. The performance of DSV-CP model is compared with the exiting DM-BD and Flex-Analytics methods. To evaluate the DSV-CP model, the following metrics are used.

i) Classification accuracy
ii) Prediction rate

(a) Performance Analysis of Classification Accuracy: Classification accuracy measures the number of correct classifications regarding the total number of big data instances in training dataset classified. The classification accuracy 'Ai ' of an individual instance in training dataset 'i' depends on a number of data correctly classified and it is measured in terms of percentage as in Equation (7).

$$A_i = \frac{DCC}{n} * 100 \qquad (7)$$

where 'DCC' signifies the number of Data Correctly Classified and 'n' is the total number of data considered for evaluation for measuring the classification accuracy. When the classification accuracy is higher, the method is said to be more efficient.

	Classification Accuracy (in %)		
Size of Big Data (in GB)	**Existing DM-BD**	**Existing Flex-Analytics**	**Proposed DSV-CP**
20	61.28	66.54	74.3
40	62.87	67.85	76.12
60	64.14	69.67	77.89
80	65.32	71.71	79.12
100	67.66	72.32	82.34
120	68.94	74.85	83.78
140	70.17	76.24	85.67
160	72.39	78.64	86.12
180	74.63	79.37	88.47
200	75.67	81.21	89.36

Table 2Resultsfor ClassificationAccuracy

Table 2 represents the classification accuracy of DSV-CP model. To determine the performance of DSV-CP model, the comparison of the classification accuracy of the model with that of the two DM-BD and Flex-Analytics methods is done. From the Table 2, it is noticed that the classification accuracy of the proposed DSV-CP model is higher when compared with the existing methods. For experimental evaluation, the size of big data is considered in the range of 20 GB to 200 GB. The impact of classification accuracy using the three different methods is demonstrated in Figure 5. As in the Figure 5, the proposed DSV-CP model

provides better performance compared to DM-BD and Flex-Analytics methods.

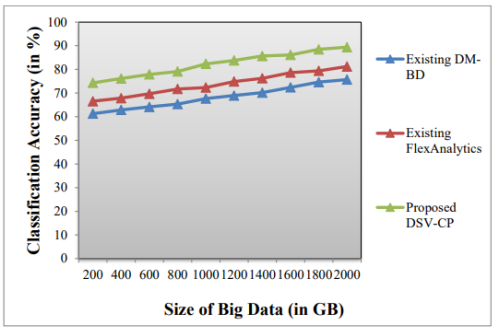

Figure 5 Performance Analysis of Classification Accuracy

Besides, with the increase in the size of big data, the classification accuracy is also increased, but comparatively, the classification accuracy using the proposed DSV-CP model is found to be higher. This is because of support vector prediction classifier in DSV-CP model which significantly reduces the misclassification errors. Due to two parallel hyperplanes applied, slack variables in the proposed model are introduced in such a way that the misclassification errors are reduced. Therefore, the classification accuracy in DSV-CP model is improved by 18.21% and 9.34% against DM-BD and Flex-Analytics methods respectively.

(B) Performance Analysis of Prediction Rate

DSV-CP model predicts the user requests based on the current data and historical facts. Prediction rate is the practice of extracting information from the existing user queries in order to share the results with the other users and to predict future outcomes and trends.

$$Prdiction\ Rate = \frac{Current\ data(size) + Historical\ Facts(size)}{Size\ of\ Big\ Data} * 100 \quad (8)$$

From Equation (8), the prediction rate is determined by summing up the current data size obtained and the size of historical facts with respect to the size of big data. When the prediction rate is high, the method is said to be more efficient.

	Prediction Rate (in %)		
Size of Big Data (in GB)	**Existing DM-BD**	**Existing Flex-Analytics**	**Proposed DSV-CP**
20	55.64	59.62	68.72
40	56.48	61.23	69.87
60	57.68	63.45	71.24
80	59.12	65.82	73.48
100	60.58	67.15	75.68
120	62.31	68.98	77.95
140	65.48	70.23	79.82
160	65.89	72.34	81.34
180	67.42	74.97	82.38
200	69.83	75.81	84.33

Table 3Resultsfor PredictionRate

The prediction rate using the three different methods, namely DSV-CP model, DM-BD and Flex-Analytics is elaborated in Table 3. The big data sizes ranging from 20 GB to 200 GB are taken into consideration for experimental purpose using Java language. From the Table 3, it is obvious that the prediction rate using the proposed DSV-CP model is higher compared to the DM-BD and Flex-Analytics methods.

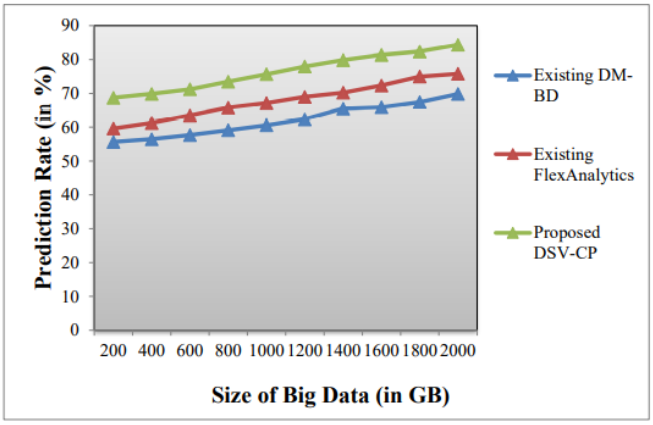

Figure 6 Performance Analysis of Prediction Rate

Figure 6 shows the prediction rate versus different sizes in the range of 20 GB to 200 GB. As shown in the Figure 6, the proposed DSV-CP model using the prediction rate provides better performance than that of the other two methods. This is also because of support vector prediction classifier that involves parallel hyperplanes to restrict hyperplanes that pass through the origin by selecting the optimal hyperplanes to classify the big data. Therefore, the prediction rate using DSV-CP is improved by 23.33% compared to DM-BD method and by 12.62% compared to Flex-Analytics method.

5. CONCLUSION AND FUTURE SCOPE:

The DSV-CP model is proposed to achieve efficient big data computation and sharing of information in cloud computing environment in this chapter. Its main goal is to improve the classification accuracy and the prediction rate of the user request information in cloud environment. DSV-CP model initially performs the data preprocessing task to efficiently remove the noise and inconsistent data in datasets which, in turn, reduces the computation time and space complexity in an effective manner. After performing the data preprocessing task, DSV-CP model uses support vector prediction classifier to effectively classify big data in cloud. The DSV-CP significantly reduces the misclassification errors by improving the search and prediction accuracy of the user request information on big data. The performance of DSV-CP model is tested with the metrics such classification accuracy and prediction rate. The experimental results show that the DSV-CP model provides better performance with an improvement in classification accuracy., LFR-CM framework is a remarkable one for executing the big data functions in parallel manner for sharing the information in cloud environment.

REFERENCES

[1] Olshannikova, E, Ometov, A, Koucheryavy, Y & Olsson, T 2015, 'Visualizing Big Data with augmented and virtual reality: challenges and research agenda', Journal of Big Data, vol. 2, no. 1, pp. 22-22.

[2] Liang, Y, Wu, D, Liu, G, Li, Y, Gao, C, Ma, ZJ & Wu, W 2016, 'Big data-enabled multiscale serviceability analysis for aging bridges☆', Digital Communications and Networks, vol. 2, no. 3, pp. 97-107.

[3] Assunção, MD, Calheiros, RN, Bianchi, S, Netto, MAS &Buyya, R 2015, 'Big Data computing and clouds: Trends and future directions', Journal of Parallel and Distributed Computing, vol. 79-80, pp. 3-15.

[4] Baro, E, Degoul, S, Beuscart, Rg&Chazard, E 2015, 'Toward a literature-driven definition of big data in healthcare', BioMed Research International, vol. 2015.

[5] Pääkkönen, P &Pakkala, D 2015, 'Reference architecture and classification of technologies, products and services for big data systems', Big Data Research, vol. 2, no. 4, pp. 166-186.

[6] Xiongpai, Q, Huiju, W, Xiaoyong, D & Shan, W 2010, 'Parallel techniques for large data analysis in a futures trading evaluation service system', in 9th International Conference on Grid and Cooperative Computing (GCC), pp. 179-184.

[7] Yun, X, Wu, G, Zhang, G, Li, K & Wang, S 2015, 'FastRAQ: A fast approach to range-aggregate queries in big data environments', IEEE Transactions on Cloud Computing, vol. 3, no. 2, pp. 206-218.

[8] Ergu, D, Kou, G, Peng, Y, Shi, Y & Shi, Y 2013, 'The analytic hierarchy process: task scheduling and resource allocation in cloud computing environment', The Journal of Supercomputing, vol. 64, no. 3, pp. 835-848.

[9] Zhou, X, Qin, X & Li, K 2015, 'Parallel techniques for large data analysis in the new version of a futures trading evaluation service', Big Data Research, vol. 2, no. 3, pp. 102-109.

[10] Reddy, VA & Reddy, GR 2015, 'Study and Analysis of Big Data in Cloud Computing', International Journal of Advance Research in Computer Science and Management Studies, vol. 3, no. 6, pp. 416-422.

[11] Triguero, I, Peralta, D, Bacardit, J, Garcia, S & Herrera, F 2015, 'MRPR: A Map Reduce solution for prototype reduction in big data classification', Neurocomputing, vol. 150, no. A, pp. 331-345.

[12] Suthaharan, S 2016, 'Machine Learning Models and Algorithms forBig Data Classification', vol. 36, pp. 17-29.

[13] Wu, X, Zhu, X, Wu, G-Q & Ding, W 2014, 'Data mining with big data', IEEE Transactions on Knowledge and Data Engineering, vol. 26, no. 1, pp. 97-107.

[14] Pawar, A & Dani, A 2016, 'A Novel Approach for Protecting Privacy in Cloud Storage based Database Applications', vol. 15, pp. 167-178.

[15] Wang, L, Tao, J, Ranjan, R, Marten, H, Streit, A, Chen, J & Chen, D 2013, 'G-Hadoop: MapReduce across distributed data centers for data-intensive computing', Future Generation Computer Systems, vol. 29, no. 3, pp. 739-750.

[16] O'Driscoll, A, Daugelaite, J & Sleator, RD 2013, "Big data', Hadoop and cloud computing in genomics', Journal of biomedical informatics, vol. 46, no. 5, pp. 774-781.

[17] Kchaou, H, Kechaou, Z &Alimi, AM 2015, 'Towards an Offloading Framework based on Big Data Analytics in Mobile Cloud Computing Environments', Procedia Computer Science, vol. 53, pp. 292-297.

[18] Zou, H, Yu, Y, Tang, W & Chen, H-wM 2014, 'FlexAnalytics : A Flexible Data Analytics Framework for Big Data Applications with I / O Performance Improvement', Big Data Research, vol. 1, pp. 4-13.

[19] Peralta, D, del Río, S, Ramírez-Gallego, S, Triguero, I, Benitez, JM & Herrera, F 2015, 'Evolutionary feature selection for big data classification: A mapreduce approach', Mathematical Problems inEngineering, vol. 2015.

[20] Barkhordari, M &Niamanesh, M 2015, 'ScaDiPaSi: An Effective Scalable and Distributable MapReduce-Based Method to Find Patient Similarity on Huge Healthcare Networks', Big Data Research, vol. 2, no. 1, pp. 19-27.

[21] Razzaghi, T, Roderick, O, Safro, I & Marko, N 2016, 'Multilevel weighted support vector machine for classification on healthcare data with missing values', PLoS ONE, vol. 11, no. 5.

[22] Inukollu, V, Arsi, S &Ravuri, Sr. 2014, 'Security Issues Associated With Big Data in Cloud Computing', International Journal of Network Security & Its Applications, vol. 6, no. 3, pp. 45-56.

[23] Bautista Villalpando, L, April, A &Abran, A 2014, 'Performance analysis model for big data applications in cloud computing', Journal of Cloud Computing: Advances, Systems and Applications, vol. 3, no. 1, pp. 19-38.

[24] Karthick, N & Agnes Kalarani, X 2015, 'An improved method for handling and extracting useful information from big data', Indian Journal of Science and Technology, vol. 8, no. 33.