International Research Journal of Advanced Engineering and Science
ISSN: 2455-9024

Privacy Preserving in Big Data using Cloud Virtulaization

K. Venkatachalapathy[1], V. S. Thiyagarajan[2], K. Ranjani[3]
[1]Faculty of Engineering and Technology, Annamalai University, Chidambaram, India.
[2, 3]Department of computer Science and Engineering, Annamalai University, Chidambaram, India.
Email address: [2]thiyagu.cse86@gmail.com

Abstract**—** *The large amount of sensor data leads to complexity of Big Data, which contains both necessary as well as unnecessary information. In order to get the necessary information from this big data, there is a need for classification and prediction techniques. Here the classification is done by two different algorithms namely, C4.5 and proposed work contains combination of C4.5 and SVM (Support Vector Machine). C4.5 is an algorithm used to generate a decision tree which is used for classification, and for this reason, it is often referred to as a statistical classifier. A Support Vector Machine (SVM) performs classification in the basis of multiclass classification, i.e., "one against many" where each category is split out and all of the other categories are merged. The performance of both classifiers is analyzed. The resultant shows proposed work performs better classification when compared to C4.5 classifier. The future predictions are also calculated and saved in the form of dataset. This dataset can be retrieved by server and partitioned as packets. Then, it is transferred to more than one client simultaneously by means of interfacing unit. Privacy preservation can also be achieved by encryption and decryption while the transfer of data takes place. This process reduces the overload of transferring entire data.*

Keywords**—** *C4.5, SVM, virtualization, partitioning, big data.*

I. INTRODUCTION

Big data is nothing but structured as well as unstructured, uncertain, real-time data that is present in a massive amount. Exploration of big data is nothing but the breaking a large amount of data into smaller parts for better understanding. Big data is produced as every person in world is connecting with internet and access social, commercial, educational and business sites for better understanding. The "Big Data" problem is defined as the process of gathering and examining complex data sets so large that it becomes tough to analyze and understand physically or by using on-hand data management applications [10]. Big data is a well-known term used to describe the exponential development and availability of data in both structured as well as unstructured form. Big data may be important to business and society as the Internet has become. Big data is so large that it's difficult to practice using traditional database and software techniques [5]. The big data additionally may characterize according to following:-

1) Volume: Information produced in huge scale by machine and human teamwork than traditional information. Case in point, Information created in call emphases, which is regarding call recording, labeling of queries, request, complaints and so forth.

2) Velocity: Online networking data streams create an wide-ranging Linux of conclusions and connections important to client relationship management. That is similar to messages, photographs on twitter consecutively Facebook and so on.

3) Variety: Conventional database application organized information, i.e. information arrangement and change gradually. In inverse of that non-conventional databases procedure displays confounding rate of progress.

4) Complexity: Information management in enormous information is extremely intricate assigned, when a lot of information which is unstructured creating from different sources. This must be connected, associated also interrelated to handle the data. Information investigation assists a great deal of processing and complex time for results and comprehension. To overcome this drawback, big information process are often performed through a programming model referred to as Map Reduce. Typical, implementation of the MapReduce paradigm requires networked hooked up storage and parallel processing [20]. So there is would like for data processing techniques so as to get the helpful info from complexness of information.

According to this categorization and also the recent development of the cloud computing market, cloud/network computing should be deliberated as associate degree vital chance for telecommunication/ICT players to considerably growing their share of the business ICT market presently diagrammatic by IT vendors and web players. Advantages of cloud computing may be thought of from the various views of players within the cloud ecosystem: service suppliers, partners, and users. To consider the cloud delivery model as a united platform to deliver IT and communication services over any network (fixed, mobile and worldwide coverage), and used by any end-user connected devices (PC, TV, Smartphone, machines, etc.). To deliver a rich set of communication facilities (voice and video calls, audio, video and web conferences, messaging, unified communications, content creation, workspace, broadcasting, etc.) according to a cloud multi-tenant consumption-based convention model and creating mash ups with internet 2.0 cooperative services for communication as a service (CaaS). To consider network services (L2-L3 property, and VPN and L4-L7 network services) as smart pipes "high-grade networks" for cloud service transportation and cloud interconnection (inter-cloud) in order to assure a secure and high performance finish-to-end Quality of service (QoS) for end users [9]. In distributed Data Store technique, a data structure composed of keys and values

K. Venkatachalapathy, V. S. Thiyagarajan, and K. Ranjani, "Privacy preserving in big data using cloud virtulaization," *International Research Journal of Advanced Engineering and Science*, Volume 1, Issue 3, pp. 85-91, 2016.

is distributively hold on in an exceedingly range of servers and skim and write happen in one server consistent with the request such by a key to come back a response [21].

Data Mining is regarding explaining the previous and guesswork the long run by suggests that of information analysis. Data mining could be a multi-disciplinary field which mixes statistics, machine learning, artificial intelligence and database tools. The significance of information mining applications is usually estimated to be terribly high. Many businesses have unbroken massive amounts of knowledge over years of method, and knowledge mining is capable to extract terribly valuable data from this data. The businesses are then ready to influence the extracted data into additional shoppers, more sales, and greater profits. This is also true within the producing and medical fields. Statistics is the science of collecting, classifying, summarizing, organizing, analyzing, and interpreting knowledge. Artificial Intelligence is that the study of computer algorithms handling the simulation of intelligent behaviors so as to perform those activities that ar unremarkably thought to need intelligence. Machine Learning is the study of computer algorithms to find out so as to develop mechanically through expertise. Database is the science and technology of gathering, storing and managing data therefore users will recover, add, update or remove such knowledge. Data storage is the science and technology of gathering, storing and managing data with advanced multi-dimensional reporting services in support of the call creating processes.

Data mining predicts the future by means of modeling. Predictive modeling is the process by which a model is made to predict an outcome. If the outcome is categorical it is referred to as classification and if the outcome is numerical it is so-called regression. Descriptive modeling or clustering is the duty of observations into clusters so that examinations in the same cluster are similar. Finally, association rules can find interesting associations amongst observations.

Classification is a data mining function that allots items in a collection to target groups or classes [13]. The goal of classification is to exactly predict the target class for each case in the data. Data classification is a method of data analysis that can be used to extract models describing significant data classes. [1] A classification task begins with the data set in which the class obligations are known. The simplest type of classification problem is binary classification. In binary classification, the target attribute has only two probable values: for example, high or low. Multiclass targets have more than two probable values: for example high, medium, low or unknown. Classification based techniques can be divided into two phases 1) Training phase and 2) Testing phase [2]. In the model build (training) process, a classification algorithm finds interactions between the values of the predictors and the values of the target. Classification models are tested by comparing the predicted values to known target values in a set of test data [12].

As the volume of digital information rises, there arises the necessity for more practical tools to raised find, filter and manage these resources. Therefore, developing fast and extremely correct algorithms to impromptu classify information has become an necessary a part of the machine learning and data discovery analysis. A Support Vector Machine (SVM) accomplishes classification by finding the hyperplane that maximizes the margin between the two classes. The vectors (cases) that define the hyperplane are the support vectors. The outstanding speed enhancement and the demand for fewer memory with the net learning setting modify the SVMs to be applicable to terribly large information sets. The fast on-line Support Vector Machine (SVM) classifier algorithmic program that reserves the extremely competitive classification accuracy rates of the progressive SVM solvers whereas requiring less machine resources [18]. Several methods have been proposed where typically a multi-class classifier is built by combining several binary classifiers [7], [16]. Some authors also proposed methods that consider all classes at once. Local SVM is a classification method that combines instance-based learning and statistical machine learning. It builds an SVM on the feature space region of the query point in the training set and uses it to predict its class [8].

C4.5 is an algorithmic program used to manufacture a choice tree developed by Ross Quinlan C4.5 is an extension of Quinlan's earlier ID3 algorithmic program. The decision trees created by C4.5 will be used for classification, and for this reason, C4.5 is usually cited as a applied math classifier. At each node of the tree, C4.5 selects the attribute of the information that almost all effectively splits its set of samples into subsets enriched in one category or the opposite. The splitting commonplace is the normalized data gain (difference in entropy). The attribute with the highest normalized information gain is taken to create the choice. The C4.5 algorithmic program then repeats on the smaller sublists.

II. LITERATURE SURVEY

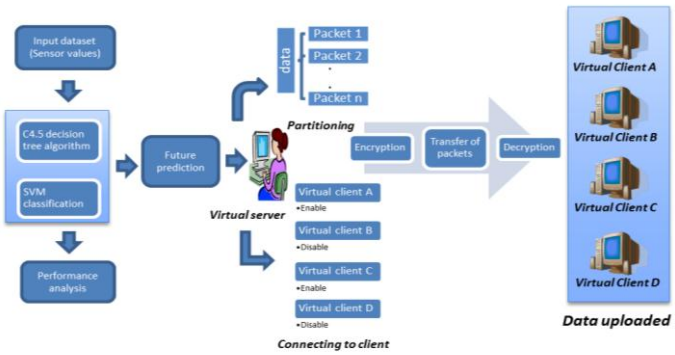
Literature presents various algorithms for effectively handling resource in BigData with Cloud environment. Here, we review some of the works presented for that. Arti Mohanpurkar et al., [3] described Balanced Partition technique which offers better performance with the help of PIG and creates a histogram for the respective partition. Vijey Thayananthan et al., [6] quantum cryptography provides maximum protection with less complication that increases the storage capacity and security strength of the big data. In this section, we need to recall the use of symmetric key with a block cipher which is suitable to control the big data security because the design of the block cipher for the big data is very simple. Chih-Wei Hsu, et al., [7] describes model of multiclass Support Vector Method based on binary classifications: "one-against-all," "one-against-one," and DAGSVM. Focus Group on Cloud Computing., [9] describes services in network (L2-L3 connectivity, and VPN and L4-L7 network services) like smart pipes "high-grade networks" for cloud transport service and interconnection of cloud (inter-cloud) in order to declarate a secure and high performance end-to-end Quality of service (QoS) for end users. Gaurav L. Agrawal et al., [11] tells about the improvement of C4.5 from ID3 algorithm like handling continuous attribute, missing attribute value, pruning tree after creation. Jeffrey Dean et al., [14] describes the run-time

K. Venkatachalapathy, V. S. Thiyagarajan, and K. Ranjani, "Privacy preserving in big data using cloud virtulaization," *International Research Journal of Advanced Engineering and Science*, Volume 1, Issue 3, pp. 85-91, 2016.

International Research Journal of Advanced Engineering and Science
ISSN: 2455-9024

system which takes care of partitioning the input data, planning the execution of program's across a set of machines, handling failures of machine, and managing the required inter-machine communication. This allows programmers without any experience with parallel and distributed systems to easily utilize the resources of a large distributed system. Lizhe Wang et, al.,[15] Commercial and public data centers offer storage, computing and software resources as cloud services, which are enabled by virtualized software/middleware stacks. Private data centers normally build basic infrastructure facilities by combining available software tools and services. They are enabled for resource sharing with grid computing middleware. This software includes cluster management system, resource management and data management system. Mahesh Pal [16] describes that the paper compares the performance of six multi-class approaches to solve classification problem with remote sensing data in term of classification accuracy and computational cost. One vs. one, one vs. rest, Directed Acyclic Graph (DAG), and Error Corrected Output Coding (ECOC) based multiclass approaches creates many binary classifiers and combines their results to determine the class label of a test pixel. Another category of multi class approach modify the binary class objective function and allows simultaneous computation of multiclass classification by solving a single optimization problem. Dr. Siddaraju et, al., [20] describes big knowledge process is achieved through a program style paradigm called MapReduce. Typical, implementation of the MapReduce paradigm requires network connected storage and parallel process.

III. System Overview



Here sensor dataset is uploaded to server page for classification process and two different classification processes should be done. First the dataset is classified by C4.5 decision tree classifier. Then the same dataset is subjected to classify according to our proposed work (SVM+C4.5). The classes and counts for each attribute value are displayed in their corresponding text boxes. The successive ratio, failure ratio, standard deviation and gain can be calculated for both the algorithm. The classification data for unique class depends up on particular attribute is shown separately. i.e., "one against many" where each category is split out and all of the other categories are merged. Performance analysis are shown in the form of graph and future predictions are calculated for each class and stored in the form of text file.

This file is retrieved by virtual server, it should be partitioned and created as packets. Then the virtual clients are selected and connected by means of interfacing unit with the server. The file which needs to transfer is encrypted and transferred as packets to more than one client simultaneously. At the client end, decryption is done automatically and original contents will be seen.

IV. Methodology

A. C4.5 Decision Tree

C4.5 is an algorithm used to produce a decision tree developed by Ross Quinlan. C4.5 is an extension lead of Quinlan's earlier ID3 algorithm. The decision trees created by C4.5 can be used for classification, and for this reason, C4.5 is often mentioned to as a statistical classifier The algorithm analyses the coaching set and forms a classifier that should be capable to properly classify each coaching and check examples. A test example is Associate in Nursing input entity Associate in Nursing the rule should predict an output price.

The input and output necessities and a pseudo code for the C4.5 rule is conferred below:

The input for the algorithm contains a set S of examples outlined by continuous or distinct attributes, each example happiness to one category. The output is a decision tree or/and a collection of rules that allots a category to a brand new case.

Pseudocode:

1. Check for the base cases
2. For each attribute j
   2.1 Find the normalized information gain ratio from splitting on attribute j
3. Let j_best be the attribute with the highest normalized information gain
4. Create a decision node that splits on j_best
5. Repeat on the sublists obtained by splitting on j_best, and add those nodes as children of node.

This algorithm has a restricted base case.

- All the samples in the list belongs to constant class. When this happens, it simply creates a leaf node for the call tree oral communication to decide on that category.
- None of the features specify any info gain. In this case, C4.5 makes a call node above the tree exploitation the certain price of the category.
- Instance of previously unseen category met. Again, C4.5 generates a call node above the tree exploitation the mean value.

The expected information needed to classify a given sample is given by equation

Gain:

The gain value can be calculated using the formula:

$$G(S,J) = E(S) - \sum_{i=1}^{m} Pr(J_i)E(S_{Ji}) \quad (1)$$

Where, $E(S) = \sum_{i=1}^{n} -Pr(C_i) * \log_2 \Pr(C_i)$ (2)

E(S) – information entropy of S.
G(S,J) – gain of S after a split on attribute J.
n – nr of classes in S.

K. Venkatachalapathy, V. S. Thiyagarajan, and K. Ranjani, "Privacy preserving in big data using cloud virtulaization," *International Research Journal of Advanced Engineering and Science*, Volume 1, Issue 3, pp. 85-91, 2016.

$Pr(C_i)$ – frequency of class C_i in S.
m – nr of values of attribute A in S.
$Pr(J_i)$ – frequency of cases that have J_i value in S.
$E(S_{Ji})$ – subset of S with items that have J_i value.

The advantages of the C4.5 are:

- Constructs a model that can be merely understood.
- Easy to apply.
- Can use both the categorical and continuous values.
- Deals with noise.

The disadvantages are:

- Small difference in information will lead to completely different call trees (especially once the variables square measure near one another in value).
- Does not work very well on a little coaching set.

C4.5 is used in classification issues and it's the foremost wide used rule for building DT. It is appropriate for world issues because it deals with numeric attributes and missing values. The algorithm will be used for making smaller or larger, more precise call trees and the rule is kind of time effective.

The C4.5 rule improves the ID3 rule by permitting numerical attributes, permitting missing values and activity tree pruning [11].

B. SVM (Support Vector Machine)

As the volume of digital information will increase, there rises the need for simpler tools to raised notice, filter and manage these resources. Therefore, emerging quick and extremely correct algorithms to mechanically classify information has become a important a part of the machine learning and information discovery analysis. The Support Vector Machine (SVM) is a supervised learning technique for information analysis, Pattern recognition, Classification and Regression analysis. It is a classification technique supported statistical learning theory, the SVM, a assuring new methodology for the classification of each linear and nonlinear information. A SVM performs classification by constructing an N-dimensional hyperplane that optimally divides the information into 2 classes. The aim of an SVM is to isolate information instances into 2 categories victimization examples of every from the coaching information to outline the ripping hyperplane. The SVM method provides AN optimally separating hyperplane in the sense that the margin among 2 teams is maximized. The subset of information instances that really outline the hyperplane square measure referred because the "support vectors", and also the margin is defined because the distance between the hyperplane and the nearest support vector. The prettiness of SVM is that if the information is linearly severable, there is a novel global minimum price. An ideal SVM analysis ought to produce a hyperplane that utterly splits the vectors (cases) into 2 non-overlapping categories. However, perfect split-up could not be attainable, or it may end in a model with such a lot of cases that the model doesn't categorise properly. In this situation SVM finds the hyperplane that maximizes the margin and reduces the misclassifications.

The idea of using a hyperplane to isolate the feature vectors into 2 groups works well but for classification with more than 2 categories, two most common methods can be used:

- "one against many" where each category is splitting out and all of the other categories are merged.
- "One against many" where k(k-1)/2 models are built where k is number of categories.

Pseudocode

- Define an optimum hyperplane: make best use of margin.
- Enlarge the above definition for non-linearly separable problems: have a penalty term for misclassifications.
- Represent data to high dimensional space where it is easier to categorize with linear decision surfaces: reformulate problem so that data is plotted implicitly to this space.

C. Proposed Work

Support Vector Machine outperforms the remaining two classifiers (Naive Bayes, C4.5) and proves to be the best among the three. SVM may have some disadvantages but that can be developed by combining SVM with other algorithms [19]. classified instances C4.5 is one of the most common algorithms for rule base classification. There are many realistic features in this algorithm such as continuous number categorization, missing value handling, etc. However in many cases it takes more processing time and provides less accuracy rate for correctly [17]. So the combination of C4.5 and SVM gives better performance.

Let S be set consisting of data sample. Suppose the class label attribute has m Distinct values defining m distinct class Ci (for i =1... m). Let Si be the total number of Sample of S in class Ci. The expected information needed to classify a given sample is given by equation

$$I(S1,S2,.....,Sm) = \sum_{i=1}^{m} -(P_i * \log_2(P_i)) \quad (3)$$

Where Pi is probability that an random sample belongs to classify Ci and estimated by Si/S. Note that a log function to base 2 is used since the information is encoded in bit.

Let attribute A have v distinct value a1,....., av. Attribute A can be used to divide S into v subsets, S1,S2,....., Sv , where Sj holds those samples in S that have value aj of A. If A were choosen as the test attribute, then these subset would resembles to the branches grown-up from the node contains the set S. Let Sij be the total number of class Ci, in a subset by Sj , The entropy (expected information) based on partitioning into subset by A, is given by equation

$$E(A) = \sum_{j=1}^{v} (S1j + S2j + ... + Smj / S) * I(Sij + ... + Smj) \quad (4)$$

More exactly the information gain, Gain(S, A) of an attribute A, relative collection of examples S, is given by equation.

$$Gain(A) = I(S1,S2,....,Sm) - E(A) \quad (5)$$

In other words gain (A) is the expected decrease in entropy caused by knowing the Value of attribute A. The algorithm computes the information gain of each attribute.

The gain ratio is defined as

Gain Ratio(A) = Gain(A)/Split Info(A)

K. Venkatachalapathy, V. S. Thiyagarajan, and K. Ranjani, "Privacy preserving in big data using cloud virtulaization," *International Research Journal of Advanced Engineering and Science*, Volume 1, Issue 3, pp. 85-91, 2016.

The attribute with the maximum gain ratio is selected as the splitting attribute.

The threshold values are increased and, the information and gain is calculated for each and every attribute, which is present in the dataset when compare to existing C4.5. Splitting the attributes is based on SVM. The splitting is based on multiclass SVM classification i.e., "one against many" where each category is split out and all of the other categories are merged.

- Successive ratio for 'n' attributes is calculated by

$$SR(val) = \sum_{i=1}^{n}((val * (\frac{i}{2})) * \log(i) * \log(i)) * n / \log(n) \quad (7)$$

Where *val=(TSR+TGain)/2* (8)

TSR= sum of information gain from all attributes.

TGain=sum of gain of all attributes.

- Failure ratio for 'n' attributes is calculated by

$$FR(val) = \sum_{i=1}^{n}(val * (\frac{i}{2})) * n / \log(n) \quad (9)$$

Where *val=(TFR+TGain)/2* (10)

TFR= sum of information gain from all attributes.

TGain=sum of gain of all attributes.

- Standard Deviation is calculated by

SD=TGain/TSR (11)

It is a measure that's wont to quantify the quantity of variation or dispersion of a group of information values. A standard deviation near zero indicates that the information points tend to be terribly near the mean (also known as the expected value) of the set, while a high normal deviation indicates that the information points ar displayed over a wider vary of values.

D. Prediction

Large amounts of sensor data have to be "interpreted" to obtain knowledge about tasks that happen in the environment Patterns in the data can be used to forecast future actions. Prediction attempts to form patterns that allow it to predict the next event(s) given the available input data.

Objective

- Anticipate inhabitant activities.
- Discover unusual occurrences (anomalies).
- Predict the right sequence of actions.
- Provide information for decision making.

Classification based prediction

- Input: Condition of the environment.
  - Attributes of the present state and previous states.
- Output: Concept description.
  - Concept specifies next event
- Prediction has to be applicable to future examples.

Prediction is vital in smart environments

  - Catches repetitive patterns (activities).
  - Helps automating events (But: only tells what will happen next; not what the system should do next).

E. Partition Algorithm

Hadoop is most popular for best execution the distributed computing, but it's simple partitioning methodology does not reserve correlation between data chunks. So there is need for partitioning Framework like FARQ in which partitioning helps for balancing data pieces into corresponding partitions. These partitions hold data for increasing the processing speed. According to big data record field partitioning algorithm is splitting and examining that particular record. Also, it is allocated a record transfer data from large tables to small tables. For bursting query performance, the partitioning algorithm plays a vital role. The selection of data is necessary for the analysis of big data because the data is present in large amount. Selection has various methods, one of the most famous method is stratified sampling in which sampling takes place among independent groups and select only one sample for development and reduction of errors. This project is based on the stratified sampling partitioning algorithm. This algorithm separates space values into different groups and subdivides groups into different portions according to server space available for particular partition. Partition algorithm is expressed for data set DS as Partition (Ds) = (G, pn) = (Vi , random [1, Vrange]) Where pn is number of a partition in group G, random function is a random number in [1; Vrange] , and Vi is a Group Identifier (GI) for the group G. For initial condition GI is equal to < 0;0; 0 > then length of the group is [0; 1]. For GI is equal to <x; 0; 0 > then length of group is [2x ; 2x + 1]. For GI is equal to < x; y ; 0 > then length of group is [2x +y ; 2x +y +1]. For GI is equal to < x; y ; z > then length of group is[2x +y +z ; 2x +y +z + 1].

Algorithm steps:

- Input: Record(R), VectorSet VTS
- Output: Partition identifier PAI
- Record has to parse into diverse column families.
- Compute Group Identifier (GI) with value ranges as stated above. Get partition vector Vp from VTS with GI and set
- Vpi=<GI ; Vrange> Set the target for Partition identifier,
- Pi =<GID; random [1; V pi * Vrange]; Build sample in partitioning Pi;
- count Pi count Pi + 1;
- sum Pi sum Pi + N;
- sample Pi sum x; y ;z ;range =count Pi;
- Ri Hash (Pi; counter Pi);
- Send Record to partition Pi; return Pi;

We use mean value of aggregation that generates samples, given as Sample = S UM =count, where S U M - sum of value from aggregation, and count- number of records in current partition. P I sent to partition is generated by input record R [22].

F. Interfacing Unit

It acts as a connecting bridge between virtual server and virtual clients in order to transfer data between them. By using this interface unit, it can able to connect and send data to more than 3 clients from server. The server and client can be connected by giving their corresponding IP address, so that it gets enabled and delivering excellent end to end performance across wide-area [5]. Port numbers for VMs are assigned in coding itself. It also indicates whether the VMs are enabled or not to send the data.

Virtualization

K. Venkatachalapathy, V. S. Thiyagarajan, and K. Ranjani, "Privacy preserving in big data using cloud virtulaization," *International Research Journal of Advanced Engineering and Science*, Volume 1, Issue 3, pp. 85-91, 2016.

International Research Journal of Advanced Engineering and Science
ISSN: 2455-9024

Virtualization is decreasing the need for physical hardware systems; saves cost and provide incremental scalability of hardware resources [4]. Virtualization needs more data transfer capacity, preparing limit and storage room, when contrast with customary server or desktop, if the physical equipment is going to have different running virtual machines on it. Business and open datacenters give registering, stockpiling, and programming assets as cloud administrations, which can be empowered by virtualized programming/middleware stacks. Private Datacenters typically form basic infrastructure services by combining available software tools and services [15].

Virtual machine

A Virtual Machine (VM) is an operating system or application environment that is installed on software which replicas dedicated hardware. The end client has the same ability on a virtual machine as they would have on committed equipment. Specific programming called a hypervisor imitates the server's CPU or PC customer, hard circle, memory, system and other equipment assets, empowering virtual machines to share the assets. The hypervisor can copy various virtual equipment stages which are disengaged from each other, permitting virtual machines to keep running on various stages, for example, Linux and Windows server working frameworks on the same hidden physical host. VMs use equipment all the more productively, which brings down the measures of equipment and related support costs. Additionally it diminishes power and cooling request. They additionally effortlessly oversaw in light of the fact that virtual equipment does not come up short. Directors can exploit virtual situations to streamline reinforcements, fiasco recovery, new positions and essential framework organization undertakings. VMs can without much of a stretch move, replicated and reallocated between host servers to conform equipment asset use. Since VMs on a physical host can devour unequal asset amounts (one may hoard the accessible physical stockpiling while another stores little), IT experts must adjust VMs with accessible assets. In this project, VMs are generated by using java coding.

G. Aggregation

The collection inquiry is only the total capacities utilized as a part of the question dialects like SQL, prophet, MySql and Sybase. There is Online Aggregate (OLA) that is utilized for enhancing the intuitive execution of database. For the compelling operations on database, clump mode is performing a key part. The customary way is that client questions and holds up till the database arrives at an end of preparing whole inquiry. On negate to OLA, the client gets expected results next to each other as question is let go. In 1997, Arti Mohanpurkar [3] incorporates that Hellerstein proposed the OLA for gathering by total questions for only one table. Here aggregation is used to grouping the packets, when it reaches the client side.

V. Performance Evaluation

The performance evaluation shows comparison between the C4.5 and the proposed work of C4.5+SVM. In the proposed work the threshold values are increased to calculate the amount of information, entropy and gain based on C4.5 and the classification is based on SVM. For each and every attribute, information, entropy and gain are calculated for both existing C4.5 and proposed work (SVM+C4.5).

The overall gain, standard deviation, successive ratio, failure ratio are calculated. Table I shows the algorithm and its corresponding values of information, entropy and gain. Figure 1 represents the chart representation corresponding to the above table.

TABLE I. Comparison of C4.5 and proposed work.

Algorithms	Successive Ratio (%)	Failure Ratio (%)	Standard Deviation (%)	Gain (%)
C4.5	18.29598	8.798507	0.819417	11.36262
SVM+C4.5	23.15639	10.71044	0.678065	18.80506

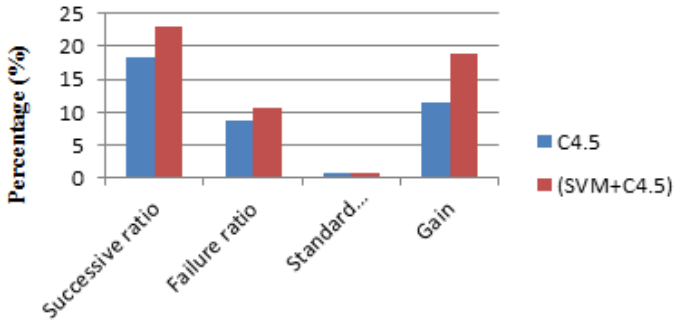


Fig. 1. Comparison of c4.5 and proposed work (svm+c4.5).

The successive ratio of proposed work is 23.156% which is more than C4.5's successive ratio 18.295%. The proposed work also having high gain of 18.805%, when compare to C4.5's gain 11.362%. A standard deviation near zero indicates that the information points tend to be terribly near the mean (also known as the expected value) of the set, while a high normal deviation indicates that the information points ar displayed over a wider vary of values. Here the Standard deviation of proposed work 0.678% is close to 0, when compare to C4.5's Standard deviation 0.819%. Even though the failure ratio of proposed work is high 10.71% when compare to C4.5's failure ratio 8.798%, the overall performance shows that proposed work is better when compare to that of existing C4.5.

VI. Conclusion

Large amount of sensor value forms the big data, which contains both useful and irrelevant information. In order to avoid transferring whole data, the relevant data can be finding out by classification and prediction techniques. The classification is done by two algorithms namely c4.5 and proposed work (SVM+C4.5). The performance of both classifiers is analyzed by means of standard deviation, gain, success ratio and failure ratio. The resultant shows, proposed work performs better classification when compared to c4.5 classifier. Also, the future predictions are calculated. By the classification and prediction, only the predicted data tends to transfer to the client. It reduces the overload of transferring entire data. Using big data sensor values, the needed values

K. Venkatachalapathy, V. S. Thiyagarajan, and K. Ranjani, "Privacy preserving in big data using cloud virtulaization," *International Research Journal of Advanced Engineering and Science*, Volume 1, Issue 3, pp. 85-91, 2016.

are predicted and this can be transferred from virtual server to more than one virtual client by means of interfacing unit in the form of packets with privacy preservation of data.

In future, the real time dataset, which can be collected dynamically from satellite, sensors, social media, etc., can be used instead of this static data set. The online aggregate queries are also used to fetch the necessary data.

REFERENCES

[1] H. Chauhan and A. Chauhan, "Implementation of decision tree algorithm c4.5," *International Journal of Scientific and Research Publications*, vol. 3, no. 10, 2013.

[2] A. Prabakar Muniyandi, R. Rajeswari and R. Rajaram, "Network anomaly detection by cascading K-Means clustering and C4.5 decision tree algorithm," *International Conference on Communication Technology and System Design*, vol. 30, pp. 174-182, 2012.

[3] A. Mohanpurkar and P. kumar Kale, "Big data analysis using partition technique," *International Journal of Computer Science and Information Technologies(IJCSIT)*, Vol. 6, issue 3, pp. 2871-2875, 2015.

[4] A. C. Arpaci Dusseau, R. H. Arpaci Dusseau, D. E. Culler, J. M. Hellerstein, and D. A. Patterson, "High-performance sorting on networks of workstations," ACM 0-89791 -911 -419710005, pp. 243-254, 1997.

[5] A. S. Pillai and J. J. Panackal, "Adaptive utility-based anonymization model: Performance evaluation on big data sets," *Procedia Computer Science*, vol. 50, pp. 347-352, 2015.

[6] V. Thayananthan and A. Albeshri, "Big data security issues based on quantum cryptography and privacy with authentication for mobile data center," *Procedia Computer Science*, vol. 50, pp. 149-156, 2015.

[7] C.-W. Hsu and C.-J. Lin, "A comparison of methods for multi-class support vector machines," *IEEE Transactions on Neural Networks*, vol. 13, no. 2, pp.415-425, 2002.

[8] N. Segata and E. Blanzieri, "Fast local support vector machines for large datasets," *Springer*, pp. 295–310, 2009.

[9] Focus Group on Cloud Computing, "Cloud computing benefits from telecommunication and ICT perspectives," part. 7, 2012.

[10] F. Zimmerman, "Bringing big data into the enterprise," *Enterprise Executive Magazine*, 2013.

[11] G. L. Agrawal and H. Gupta, "Optimization of C4.5 decision tree algorithm for data mining application," *International Journal of Emerging Technology and Advanced Engineering*, vol. 3, no. 3, 2013.

[12] H. Kaur and H. Kaur, "Proposed work for classification and selection of best saving service for banking using decision tree algorithms," *International Journal of Advanced Research in Computer Science and Software Engineering*, vol. 3, no. 9, 2013.

[13] S. Sharma, J. Agrawal, and S. Sharma, "Classification through machine learning technique: C4.5 algorithm based on various entropies," *International Journal of Computer Applications*, vol. 82, no. 16, 2013.

[14] J. Dean and S. Ghemawat, "MapReduce: Simplifed data processing on large clusters," Google Inc, OSDI 2004.

[15] L. Wang and R. Ranjan, "Processing distributed internet of things data in clouds," *IEEE Cloud Computing*, pp. 2325-6095, 2015.

[16] M. Pal, "Multiclass approaches for support vector machine based land cover classification," 2008.

[17] M. M. Mazid, A. B. M. Shawkat Ali, and K. S. Tickle, "Improved C4.5 algorithm for rule based classification," *Recent Advances In Artificial Intelligence, Knowledge Engineering and Data Bases*, pp. 296-301.

[18] S. Ertekin and G. Hopper, "Efficient support vector learning for large datasets," 2006.

[19] M. Trivedi, S. Sharma, N. Soni, and S. Nair, "Comparison of text classification algorithms," *International Journal of Engineering Research & Technology (IJERT)*, vol.4, no. 02, 2015.

[20] Siddaraju, C. L. Sowmya, K. Rashmi and M. Rahul, "Efficient analysis of big data using map reduce framework," *International Journal of Recent Development in Engineering and Technology*, vol. 2, issue 6, 2014.

[21] S. Tsuchiya, Y. Sakamoto, Y. Tsuchimoto, and V. Lee, "Big data processing in cloud environments," *FUJITSU Sci. Tech. J.*, vol. 48, no.2, pp. 159-168, 2012.

[22] K. Venkatachalapathy, V. S. Thiyagarajan, A. Ayyasamy, and K. Ranjani, "Big data with cloud virtualization for effective resource handling," *International Journal of control Theory and Applications*, vol. 9, no. 2, pp. 435-444, 2016.