Effective FrameworkwithFinite Client Transfer Data set for Weather Prediction using Data Mining Techniques

L. R. Aravind Babu[1] and Dr.K.Venkatachalapathy[2]
[1]Assistant Professor in Information Technology, FEAT, Annamalai University,
[2]Professor and Head, Division of Computer and Information Science, Annamalai University

Abstract:
The importance of big data insiststhe data sets that are so large in such a way that traditional data processing applications are not enough to process it. In order to get the necessary information from the big data, there is a need for classification technique and also we use techniques for prediction using those data sets. Here the classification is done by two different algorithms namely C5.0 and SVC (Support Vector Clustering) algorithm, where both of them are combined in proposed work to give efficient results in classification of the required data sets. C5.0 is an algorithm used to generate a decision tree which is used for classification, and for this reason it is often refer to as a statistical classifier. It performs winnowing in such a way that the decision tree becomes more accurate and removes the attributes which may be unhelpful. The SVC is a statistics clustering algorithm that does not make any presumption on the number of the clusters in the data. The performance of both classifiers was monitored and analyzed. The result of the proposed work shows better classification when compared to the single use C5.0 classifier. The future weather predictions are also been calculated and saved in the form of dataset virtualization.

Keywords**:** *C5.0; SVC algorithm; Winnowing; partitioning*

I. INTRODUCTION

Big data is nothing but structured as well as unstructured, uncertain, real-time data that is present in a massive amount. Classification of big data is nothing but the breaking a large amount of data into smaller parts for better understanding. The “Big Data” problem is defined as the process of gathering and examining complex data sets which are so large, in such a way that it becomes tough to analyses 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. 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 effective tools to better find, filter and manage these resources. Therefore, developing fast and highly accurate algorithms to spontaneously classify data has become an important part of the machine learning and knowledge discovery research. A SVC suits for low volume data and it requires a Pre-processing step for high volume data.In our proposed system we use C5.0 as a pre-processing step to the SVC algorithm where the complexity is reduced and efficiency is increased by combining both of them. SVC uses kernel function to map data points to a high dimensional area. In order to denote the precise area in data space, the Support Vector Domain Description

(SVDD) is used by SVC. A decision function is provided by SVDD to tell whether a given input is inside the feature space sphere or not. It is done so, to indicate whether particular point belongs to the support of distribution. F(x) is the decision function, where x is the value of the function. When it is inside the feature space sphere it returns a value greater than 0. Otherwise it returns negative values. C5.0 is an algorithm used to produce a decision tree. It is an extension to the ID3 algorithm. At each node of the tree, C5.0 selects the attribute of the data that most effectively splits its set of samples into subsets enriched in one class the other. The splitting standard is the normalized information gain (difference in entropy). The attribute with the highest normalized information gain is taken to make the decision. The C5.0 algorithm then repeats on the smaller sub-lists.

II.LITERATURE SURVEY

Literature presents various algorithms for effectively handling resource in Big Data with Cloud environment. Here, we review some of the works presented for that. ArtiMohanpurkar et al.,[3]described Balanced Partition technique which offers better performance with the help of PIG and creates a histogram for the respectivepartition.VijayThayananthan 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 (L4-L7 connectivity and L2-L3 network services). 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 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 centres offer storage, computing and software resources as cloud services, which are enabled by virtualized software/middleware stacks. Private data centres 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 data processing can be achieved through a program design paradigm known as MapReduce. Typical, implementation of the MapReduce paradigm requires network attached storage and parallel processing.

III. SYSTEM OVERVIEW

Here sensor dataset absorbs data from the pool of big data and it is uploaded for classification process and two different classification processes should be done. First the dataset is classified by C5.0 decision tree classifier. This file is retrieved by virtual server where 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 be transferred is encrypted and transferred as packets to more than one client simultaneously. On the other hand at the client's end, decryption is done when it receives the data packets and original contents will be seen. The successive ratio, failure ratio, standard deviation and gain can be calculated for both the algorithm and then it is compared with the graph. 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

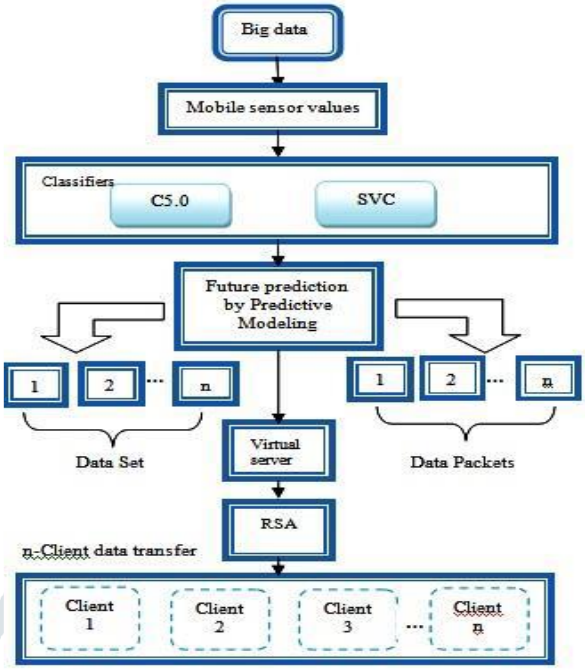

Fig.1 System Architecture

Then the same dataset is subjected to classify with Support vector clustering according to our proposed work (SVC+C5.0). The classes and counts for each attribute value are displayed in their respective text boxes.

IV.METHODOLOGY

4.1 C5.0 Decision Tree

C5.0 is an algorithm used to produce a decision tree. C5.0 is an extension lead of C4.5 algorithm. This algorithm undergoes analysis with the training set and creates a classifier that must be capable to accurately classify the training data sets. Some recent data mining applications are characterized by very high dimensionality, with plethora attributes. C5.0 can automatically winnow those attributes before a classifier is constructed, discarding those that appear to be only least relevant. For high-dimensional applications, winnowing can lead to smaller classifiers and higher predictive accuracy, and can often reduce the time required to generate rule sets. Further, C5.0 provides facilities for defining new attributes as functions of other attributes.

The advantages of the C5.0 are:

- Constructs a model that can be simply interpreted.
- Several new data types in addition to C4.5
- Supports sampling and cross validation.
- Minimize expected misclassification cost.
- More memory efficient than C.4.5

The disadvantages are:

- Minute difference in data can lead to different variation in the decision trees.
- When the decision tree is smaller the results are similar to previous version with no much improvement.

C5.0 is a frequently used algorithm for building decision tree. The algorithm can be used for creating various size of decision tree and also precise a decision trees which are always time effective. The C5.0 algorithm is advancement to the C4.5 algorithm by having several advanced attributes such as tree pruning, allowing missing values[11].

4.2 SVC (Support Vector Clustering)

There is a plethora of information which is increasing day by day; simultaneously the need for classification of that information into necessary data is a vital part in big data classification. So the classification algorithms which are more effective along with less time consumption are required.The Support Vector Clustering (SVC) is

a supervised learning technique for Data analysis, Pattern recognition, and Classification and Regression analysis. SVC is a similar method that also builds on kernel functions and is appropriate for unsupervised learning and data-mining. Clustering may proceed according to some model and also it follows a pattern by grouping points according to some attributes of those points and also similarity measure may be done as like that of hierarchical clustering. The boundaries for cluster can be formed by considering the area in the data space where there is low number of data. The SVC takes this path to form the cluster and its boundaries. When the input examples are denser or concentrated, Support Vector Domain Description (SVDD) is used by SVC to demarcate the region in data space. Clusters are identified or classified through SVC by identifying pair of data points from heterogeneous clusters and also considering the line segment which has to connect the data points must pass through a particular area in the data space. For the set of points **xi** and **xj**the *i*, *j* element of *A* is given by,

Aij= {1, if *f*(**x**)>0 for **x**, where line segment bridging **xi** and **xj.**

{0, otherwise.

V PROPOSED WORK

SVC may have some disadvantages but that can be developed by combining SVC with other algorithms[19]. Classified instances C5.0 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 C5.0 and SVC 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) = Σ ……Eq(3).,w**here 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) = Σ ……Eq(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) ……Eq.(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).

The attribute with the maximum gain ratio is selected as the splitting attribute. The threshold values are increased. For every particular attribute the information about that attribute and the gain percentage is calculated, which is present in the dataset when compare to existing C5.0. 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 Standard Deviation is calculated by

SD=TGain/TSR ……Eq(11)

When the data points are close to the mean, then the standard deviation is approximately near to 0. On the other hand when the data points have higher range of values then the standard deviation leads to higher values.

5.1Prediction

Large amounts of sensor data have to be "interpreted" to obtain knowledge about tasks that occur in a dynamic environment. The future actions could be forecasted by using patterns of those data. Prediction attempts to form patterns that allow it to predict the next event in accordance to the given available input data.

OBJECTIVE

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

5.2 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 an important role. The selection of data is necessary for the analysis of big data because the data is present in a huge 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 values into various groups and subdivides groups into different portions according to space available. 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].

5.3 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 performs end to end

communication over a large coverage 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.

5.4 Virtualization

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 servers having different virtual machines. Business and open data centres give registering, stockpiling, and programming assets as cloud administrations, which can be empowered by virtualized programming/middleware stacks. Private Data centres typically form basic infrastructure services by combining available software tools and services [15].

5.5 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, ArtiMohanpurkar[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.

VI. PERFORMANCE EVALUATION

The performance evaluation shows comparison between the C5.0 and the proposed work of C5.0+SVC. In the proposed work the threshold values are increased to calculate the amount of information, entropy and gain based on C5.0 and the classification is based on SVM. For each and every attribute, information, entropy and gain are calculated for both existing C5.0 and proposed work (SVC+C5.0).

Table 1: Comparison of C4.5 and Proposed work (C5.0+SVC)

Algorithm	Success Ratio	Failure Ratio	Standard Deviation	Gain
C5.0	18.275	8.7585	0.818	11.35
SVC	23.146	10.710	0.688	18.80
C4.5	12.24	10.8	0.75	9.02

The overall gain, standard deviation, successive ratio, failure ratio are calculated. Table 1 shows the algorithm and its corresponding values of information, entropy and gain. Figure 2 represents the chart representation corresponding to the above table. The successive ratio of proposed work is 23.146% which is more than C5.0‟s successive ratio 18.275%. The proposed work also having high gain of 18.801%, when compare to C5.0‟s gain 11.352%. Here the Standard deviation of proposed work 0.688% is close to 0, when compare to C5.0‟s Standard deviation 0.818%. Even though the failure ratio of proposed work is high 10.71% when compare to C5.0‟s failure ratio 8.758%, the overall performance shows that proposed work is better when compare to that of existing C5.0.

VII.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 C5.0 and proposed work (SVC+C5.0). 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 C5.0 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 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.

VIII.REFERENCES

[1]HarvinderChauhan and AnuChauhan, "Implementation of decision tree algorithm C5.0", International Journal of Scientific and Research Publications, Vol. 3, No. 10, October 2013.
[2]AmuthanPrabakarMuniyandi, R. Rajeswari and R. Rajaram, "Network Anomaly Detection by Cascading K-Means Clustering and C5.0 Decision Tree algorithm", International Conference on Communication Technology and System Design, vol. 30, pp.174-182, 2012.
[3] ArtiMohanpurkar and Prasad 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]Andrea C. ArpaciDusseau, Remzi H. ArpaciDusseau, David E. Culler, Joseph M. Hellerstein and David A. Patterson, "High-Performance Sorting on Networks of Workstations", ACM 0-89791 -911 -419710005, pp.243-254, 1997.
[5] Anitha S Pillai and Jisha Jose Panackal, "Adaptive Utility-based Anonymization Model: Performance Evaluation on Big Data Sets", Procedia Computer Science, Vol. 50, pp. 347-352, 2015.
[6]VijeyThayananthan and AiiadAlbeshri, "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]Chih-Wei Hsu and Chih-Jen Lin, "A Comparison of Methods for Multi-class Support Vector Machines", IEEE Transactions on Neural Networks, Vol. 13, No. 2, pp.415-425, August 7, 2002.
[8]Nicola Segata and Enrico 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, February 2012.
[10]Fred Zimmerman., "Bringing Big Data into the Enterprise", Enterprise Executive Magazine, June 2013.
[11]Gaurav L. Agrawal and Prof. Hitesh Gupta, "Optimization of C5.0 Decision Tree Algorithm for Data Mining Application", International Journal of Emerging Technology and Advanced Engineering, vol. 3, No. 3, March 2013.