Fast and Efficient Agro Data Classification Model for Agriculture Management System using Hierarchical Cloud Computing

Kuldeep P. Sambrekar, Vijay S. Rajpurohit

Abstract: ***Data analytics (DA), Internet of Things (IoT) and cloud computing framework are employed to build a cost efficient and productive agriculture management system. The remote sensing forecasting and GIS Technology provide various sensory information to stake holders/users such as rainfall pattern, weather related data (such as temperature, humidity, pressure etc.). These sensory data are of unstructured format. The existing system lack efficiency in performing analysis on such data. Since it fails to bring good tradeoff between speedup and memory efficiency. To overcome these research challenges, this work presents an Accurate Classification Model (ACM) for Agriculture Management System (AMS). Firstly, a selective clustering algorithm is proposed to classify unstructured multi-dimensional selective agriculture data to structured format. Further, this work presents a novel hierarchical clustering model to perform clustering on output data of selective clustering algorithm and stores the data on standard Hierarchical cloud storage architecture. A parallel algorithm to perform classification of structured data using Hadoop MapReduce framework is presented. Experiments are conducted on real-time agricultural data. The results obtained indicate a considerable improvement over exiting model in terms of computation cost, latency, accuracy, memory efficiency and speedup.***

Keywords: Agriculture data clustering, Map-reduce framework for agriculture, Cloud data Storage optimization, Hierarchical data on cloud.

I. INTRODUCTION

Agriculture is the backbone of most developing countries such as India where about 70% people depends directly or indirectly on it and about 40% contributes to Gross National Product (GNP).Achieving good productivity aid in attaining higher GDP growth of a country [1]. For attaining better productivity, timely and accurate information of data such as type of crop grown, crop yield, crop growth condition, rainfall pattern, weather related data (such as temperature, humidity, pressure etc.) and so on is required. To collect such data sensor are placed across agriculture field and globe. The agro data sensed by these sensor are obtained though gateway or internet and then these sensory data are transmitted to cloud computing environment. With the adoption of Internet of Things (IoT) and Cloud computing framework [2] huge volume of unstructured raw agro data is continuously being collected.

Storing and performing analysis on such unstructured data on cloud platform for providing smart agro farming requires efficient mechanism [3], i.e., the model should minimize computation cost.

Performing analysis on huge agro related unstructured high dimensional data into structured form L. Kuang, L. T. Yang [4] presented a data dimension reduction and classification technique. The model considers heterogeneous platform for scalability and dimension reduction to speed-up classifying high dimensional data. However, during dimension reduction some important feature are neglected. As a result, accuracy of their model is not efficient. Generally, large collection of points are involved, resulting in requirement for fast classifying model to assure an optimal computing time. However, for high dimensional data, no exiting algorithm can offer nearest neighbour (NN) algorithm speed-up with respect to linear classification. As a result, some approximate classification algorithm, compromise accuracy for the sake of efficiency.

Recently, some researchers Y. Gong[5], T. Ge[6],L.Bao [7], and D. Cozzolino[8] aimed at addressing this tradeoff issues i.e., they either focused addressing memory or time efficiency. In [5] and [6]researchers have addressed the case of very huge high dimensional data that do not fit into memory. On the other side in [7] and [8], have addressed the issue of high dimensional data and its associated structure can fit in memory, in this case processing time becomes the critical issues, and outcomes is measured in terms of accuracy and speedup and memory utilization is compromised. To bring a good trade-off between memory and I/O in [9] presented a Reliable Order-Statistics based Approximate NN search Algorithm (ROSANNA) for classifying unstructured high dimensional agro data. However, at some instance the speedup may not validate the utilization of additional memory. Since, the memory overhead of utilizing multiple arbitrary trees increases linearly with the size of trees. Further, the exiting model are designed classifying single dimension.

This work aimed at overcoming these challenges and present Accurate Classification Management (ACM) model. Firstly, a selective clustering algorithm to classify unstructured multi-dimension high dimensional data to structure form.

Revised Manuscript Received on February 05 2019.

Kuldeep P. Sambrekar, Dept. of CSE, Gogte Institute of Technology, Belagavi, Karnataka, India,

Vijay S. Rajpurohit, Dept. of CSE, Gogte Institute of Technology, Belagavi, Karnataka, India,

Post processing of selective clustering algorithm, this work further presents a novel hierarchical clustering and perform clustering on output data of selective clustering algorithm and store or place the data on multi-level (Hierarchical) cloud storage architecture. Further, a parallel algorithm to perform classification using Hadoop MapReduce framework is presented.

The Contribution of research work is as follows

- This work presents a selective clustering algorithm to classify multi-dimension high dimensional unstructured agro data to structured form.
- Further, this work presents a novel hierarchical clustering algorithm to further classify data and store on multi-level cloud architecture.
- Parallel algorithm to perform classification using Hadoop MapReduce framework is presented.
- Experiments are conducted on real-time agriculture data.
- ACM model attain good performance i.e., it reduces computation cost, latency, total CPU time, accuracy, memory efficiency and speedup considering varied real-time scientific and data intensive application.

The rest of the paper is organized as follows. In section II the proposed accurate classification and storage management model for multi-level cloud based agriculture storage management system is presented. In penultimate section experimental study is carried out. The conclusion and future work is described in last section.

II. ACCURATE CLASSIFICATION MODEL FOR AGRICULTURE MANAGEMENT SYSTEM

This work presents a fast and Accurate Classification Model (ACM) for performing analysis on unstructured agriculture data and store them across different cloud storage level (provider). Firstly, a selective clustering algorithm is presented to classify unstructured agriculture related data into structured format. Further, a multi-level/ hierarchical clustering algorithm is presented to further perform analysis on semi-structured data and store it distributive across cloud storage location (level).Then, parallel classification model using Hadoop MapReduce framework is presented to speedup classification process for relatively large data. The architecture of ACM for Multi-level cloud storage model is shown in Fig. 1.

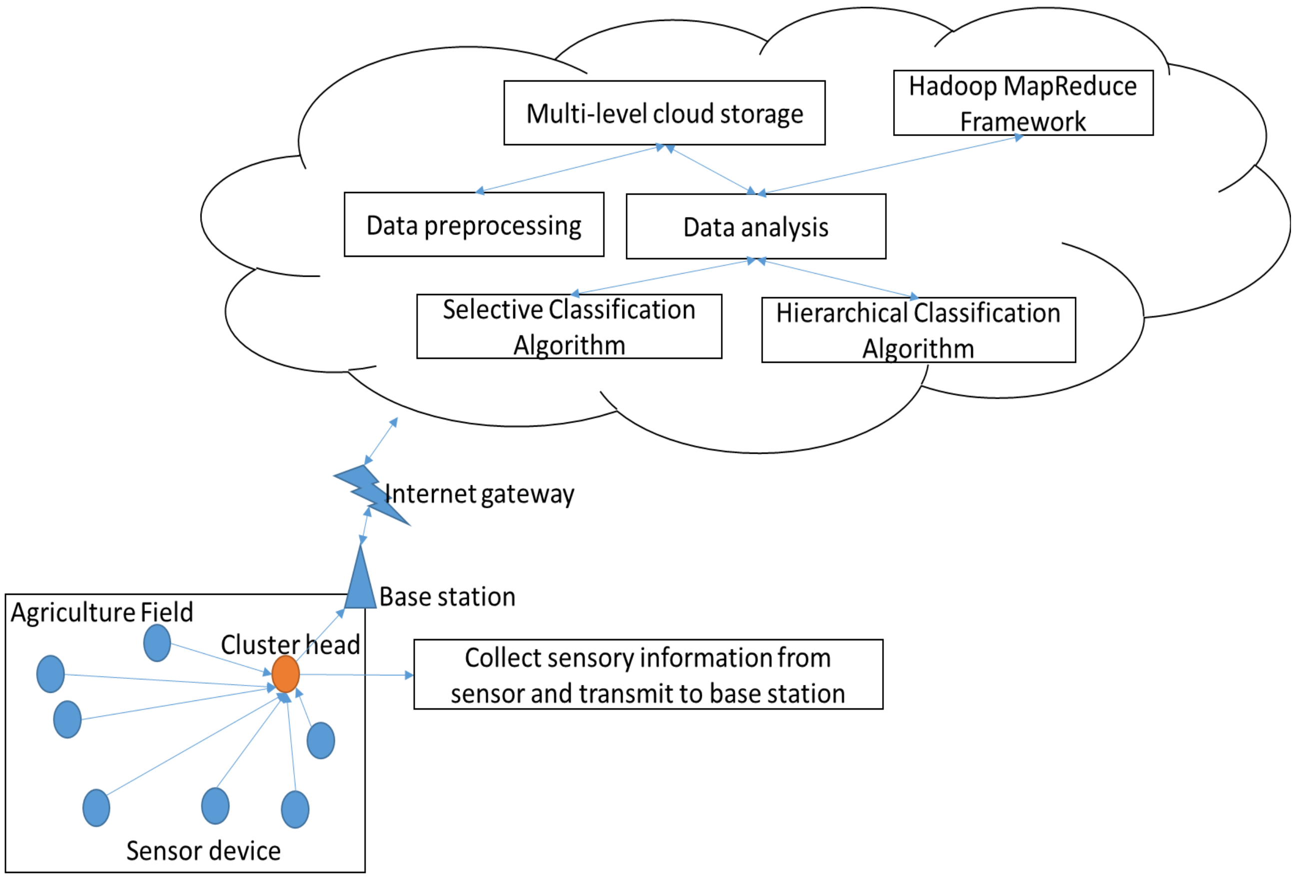

Fig. 1 Architecture of Accurate Classification Model for Multi-level cloud storage model

a) System model and dataset description

This section describes the detailed block architecture of proposed classification model as shown in Fig. 2.

Published By: Blue Eyes Intelligence Engineering & Sciences Publication

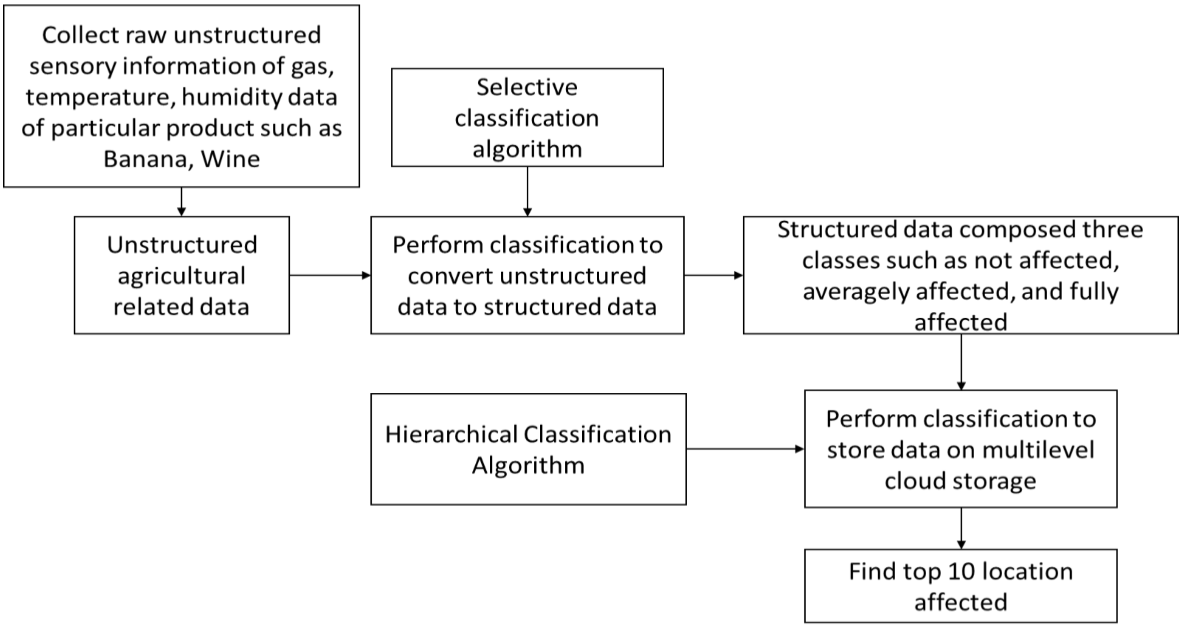

Fig. 2 Block architecture of proposed classification model

For performing analysis or classification this work used crop monitoring dataset obtained from [16]. Since no agriculture data was publically available. The data is composed of sensory data obtained from various gas, temperature, and humidity sensors. This data is used to identify the effect of gasses on wine and banana for temperature and humidity level. The data is composed of 11 attributes or dimension such as id, time, R1, R2, R3, R4, R5, R6, R7, and R8, Temperature, Humidity and is composed of 919438 data points across different location and time. More description of dataset used in this work can be obtained from [16]. For classifying these data, we applied selective clustering. The K is set to 3 (i.e., we consider three classes such as, not affected, averagely affected, and fully affected). The K can be changed based on user classification requirement.

Considering this we classify the data in to three classes and store it to cloud storage. Further we perform classification (like finding top 10 affected location) on this classified data by using multi-level or hierarchical clustering and store this classified data across different cloud storage level.

b) Clustering model for classifying unstructured raw data into structured data

The proposed selective clustering (classification) model is built by dividing the data points at each stages into L unique area using k-mean clustering. Post clustering, the same method is iteratively applied to the data points in a location area. The iterative computation is terminated when number of data points of an area is lesser than L.The proposed selective clustering model is presented in Algorithm 1.

Algorithm1: Building selective clustering algorithm
Input: Agriculture Dataset E, diverginginfluence L, maximum iteration$J_{\uparrow}$, center selection strategy to be appliedD_{str}.
Output: Selective clustering tree (Structured data).
Step 1:if$|E|L$**then**
Step 2: build terminal node with feature points in E.
Step 3:else
Step 4:Q ← choose L data points from EusingD_{str}.
Step 5: Converged←**false**
Step 6: Iterations← Zero
Step 7:while converged=**false&&** iteration< $J_{\uparrow}$**do**
Step 8:D ← cluster the feature points in E around closest centers Q
Step 9:$Q_{\mathbb{N}}$ ← averages of clusters in D
Step 10:if$Q = Q_{\mathbb{N}}$**then**
Step 11: Converged←**true**
Step 12:end if
Step 13:$Q \leftarrow Q_{\mathbb{N}}$
Step 14: iterations←iteration + 1
Step 15:end while
Step 16:for each cluster $D_j \in D$**do**
Step 17: build non-terminal node with center Q_j
Step 18: Continuously apply clustering method to the feature points in D_j
Step 19:end for
Step 20:end if

The number of cluster $L$to be considered for dividing the data at each node is a feature/attribute of the algorithm, known as the diverging influence and selecting L is significant for attaining good classification outcome.Another parameter of selective clustering algorithm is$J_{\uparrow}$, which depict the maximum iteration to perform clustering process. Considering smaller iteration aid in reducing clustering time at the cost of accuracy. However, the proposed selective clustering attain good convergence with minimal time, and lastly the parameter D_{str} is used to control the initial centers selection in clustering algorithm.

c) Clustering model for classifying semi-structured data for different level of stake holders

The multi-level clustering algorithm performs clustering operation by a disintegration of the search space by repeatedly clustering the input agricultural structured data using arbitrary data points as the centers of cluster of the non-terminal node as shown in Algorithm 2. Using algorithm 2, the data are classified based on user defined level or hierarchy and are stored on different storage level on cloud platform. The proposed multi-level clustering model is presented in Algorithm 2.

Algorithm 2: Buildingmulti-level/hierarchical clustering tree
Input: Agriculture semi-structured data E, diverging influence L, and maximum terminal size $\mathbb{T}$.
Output: Proposed hierarchical clustering tree (different storage/classification level data).
Step 1:if$E < \mathbb{T}$**then**
Step 2: build terminal node with feature points E
Step 3:else
Step 4:Q ←choose L feature points at arbitrarily from E
Step 5:D ←cluster the feature points in E around closest centers Q
Step 6:for each cluster $D_j \in D$**do**
Step 7: build non-terminal node with center Q_j
Step 8: continuously apply the clustering method to the feature points in D_j
Step 9:end for
Step 10:end if

d) Parallelizing classification using Hadoop MapReduce Framework

This work further presents parallel algorithm to perform classification using Hadoop MapReduce framework (HMR) [12]. The basic architecture of Hadoop MapReduce framework is shown in Fig. 3.

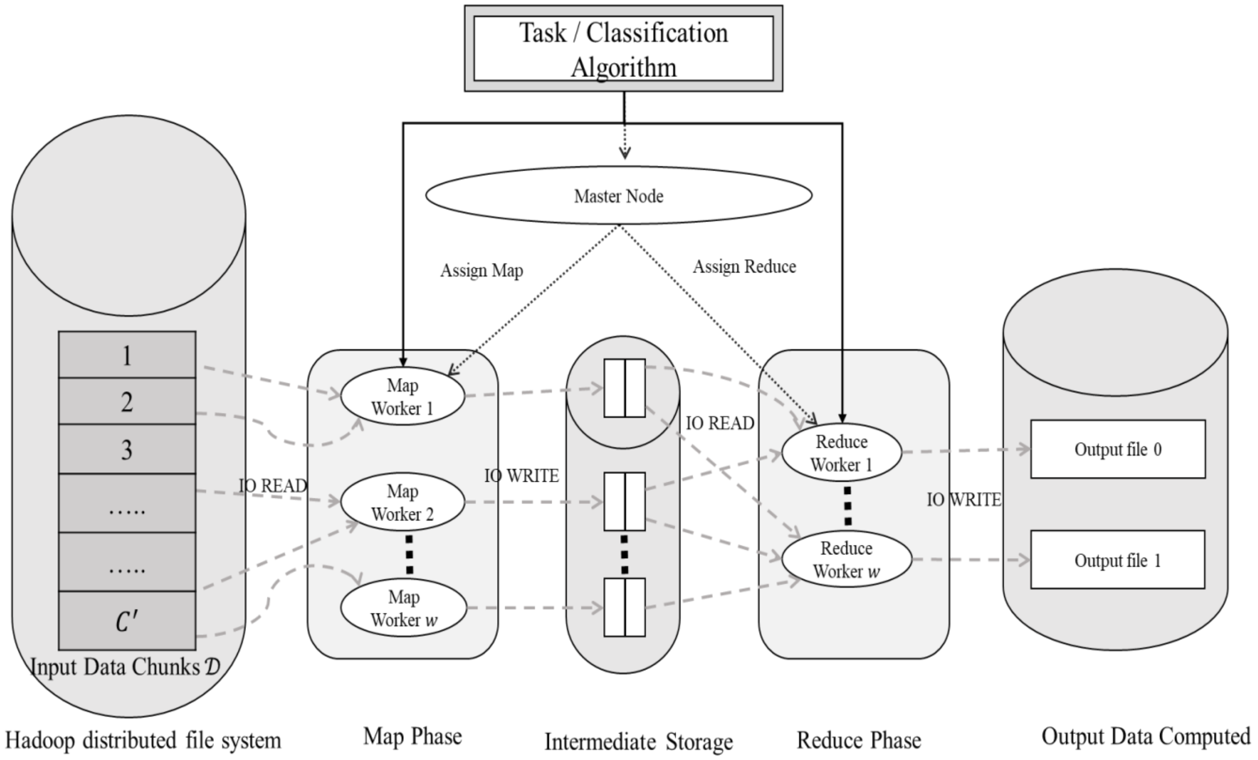

Fig. 3 The architecture of Hadoop MapReduce framework

The HMR is composed of Map and Reduce phase. In Map phase it read all input data and divide it into chunks of small data and perform execution parallel across different virtual machine. Post completion of Map Phase Reduce Phase is initialized. In this phase it reads the output of map phase and aggregate the classification output and store it in Hadoop distributed file system. Detail of Hadoop MapReduce execution can be obtained from [12]. The Algorithm to build distributed Key on Hadoop HDInsight cluster is shown in Algorithm 3.

Algorithm 3: Building distributed Key on Hadoop HDInsight cluster
Input: Data E, keyVal$Q$
Output:$ConstructKey(E, Q)$
Step 1:$j \leftarrow MR\_function()$
Step 2:E_j read chunk of the data E with respect to function j using Hadoop distributed file system.
Step 3: construct key in parallel on each worker with data E_j and keyVal Q
Step 4:$MR\_Cumulate()$ // Synchronize all workers.

This work perform classification of agriculture data using distributed architecture on multi-level cloud storage platform and our model attain good accuracy, computation time minimization, and meets real-time requirement which is experimentally shown in next section below.

III. RESULT AND ANALYSIS

This section presents performance evaluation of proposed ACM approach over exiting approach [9] in terms of CPU time, Memory overhead, accuracy and speed-up achieved considering the dataset obtained from [16]. As there is no publicly available agriculture dataset this work used crop monitoring dataset. The dataset is used to find the effect of gasses and its impact of temperature and humidity on wine and banana. Generally, the agriculture production is improved by deploying sensor device across the agriculture field. The sensors monitor the condition such as temperature, humidity etc. based on which decision like releasing water, pesticides requirement and so on. Further, the agriculture production can be enhanced by monitoring wind which aid in predicting rain arrival, cyclone etc. in particular area with less latency. So that suitable and timely decision can be taken so that minimal damage to corps is done. For that, this work compare with exiting approach [13] to evaluate the performance in terms of cost and latency incurred considering real-time scientific dataset obtained from [14] and [15] such as Inspiral. The Inspiral is utilized to identify or establish for gravitational wave signatures in data or information obtained by large-scale interferometers and is categorized by having CPU intensive tasks that requires enormous amount of memory. The experiments are conducted on windows 10 operating system, 64-bit I-7 quad core processor with 16 GB RAM with 4 GB dedicated CUDA enabled GPU. The HD Insight cluster is designed considering one master worker node and 4 slave worker node using azure HD Insight cluster using A3.Each worker node is deployed on A3 virtual machine instances which is composed of 4 virtual computing cores, 7 GB RAM and 120 GB of HDD storage space.

a) Computation cost and latency performance considering real-time data-intensive and scientific application

Experiment are conducted to evaluate the performance achieved by ACM model over exiting model [13] in terms of computation cost and latency for performing analysis on distributed computing multi-level cloud platform. Here we considered computation cost and latency performance evaluation considering Inspirral_100, workflow. The number of cloud storage node/level is varied from 20 to 80. The experiment outcome shows that the proposed ACM performs better than exiting in terms of computation cost and latency minimization. A computation cost minimization of 34.26%, 36.01%, 36.45%, and 36.74% is attained by ACM over exiting model when cloud storage/classification level size is 20, 40, 60, and 80, respectively as shown in Fig. 4. An average computation cost minimization of 35.88% is attained by ACM over exiting model considering Inspiral scientific and data-intensive workflow. A latency minimization of 16.34%, 18.57%, 19.12%, and 19.49% is attained by ACM over exiting model when cloud storage/classification level size is 20, 40, 60, and 80, respectively as shown in Fig. 5. An average latency minimization of 18.39% is attained by ACM over exiting model considering Inspiral scientific and data-intensive workflow.

*Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication*

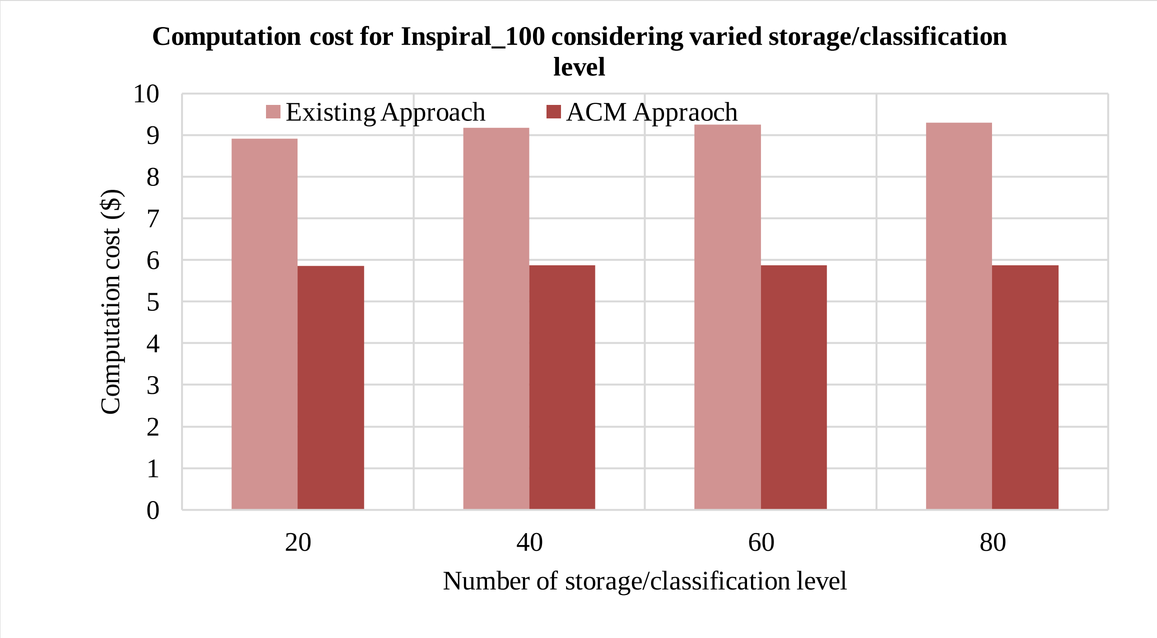

Fig. 4 Computation cost performance for varied storage/classification level for Inspiral_100

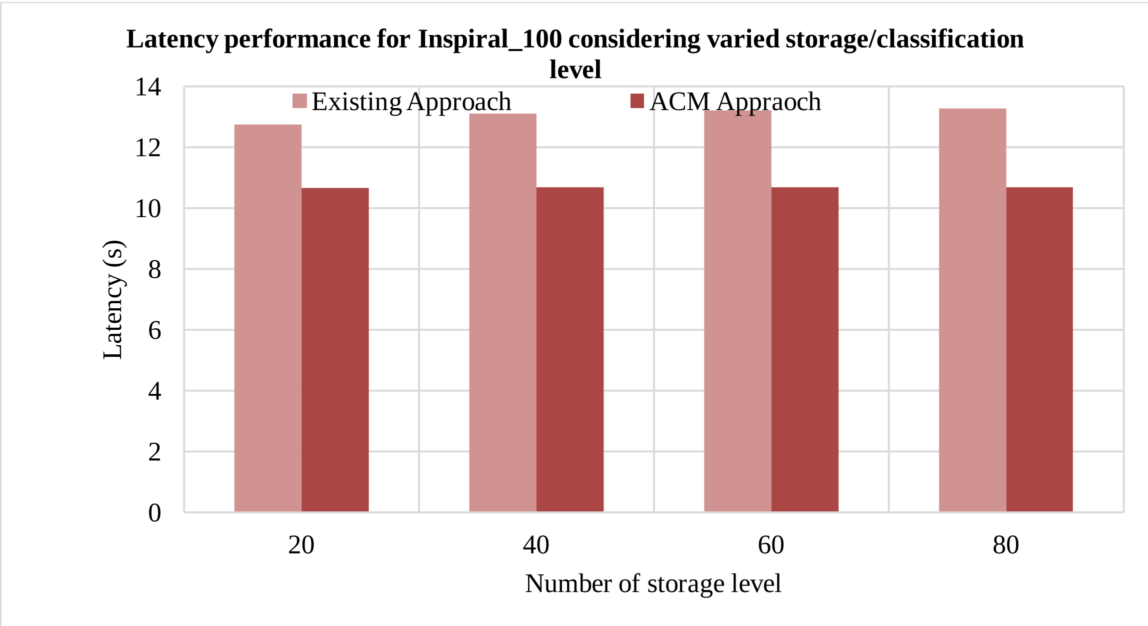

Fig. 5 Latency performance for varied storage/classification level for Inspiral_100

b) Computing time, accuracy, memory overhead and speed-up performance evaluation considering real-time dataset with varied dimension size

The overall result attained shows the proposed ACM model is efficient when compared to stat-of-art model [11], [13] in terms of minimizing computation cost and latency. Further, this work conducted experiment to evaluate the performance of ACM over existing model [9] in terms of Total CPU Time, Memory overhead, Accuracy attained in building classification tree for converting unstructured data into structured. The outcome of this evaluation is tabulated in Table I. The result shows ANN attain better performance than Random classification model. As a result, we compare proposed outcome performance improvement over ANN classification model. The ACM-Local classification model reduce total CPU time and Memory overhead by 32.85% and 55.07% respectively, and improves accuracy by 1.82%. Similarly, ACM-Hadoop classification model reduce total CPU time and Memory overhead by 95.86% and 84.05% respectively, improves accuracy by 1.82%and attain speedup of 16. Further we also evaluated the effect of dimension size on classification performance which is shown in Fig. 6. We have varied the size of dimension as 5, 7, 9, and 11 as shown in Table II and evaluated the classification outcome in terms of total CPU time, Accuracy and Memory overhead. The experiment outcome shows when dimension size is increased the computation time and memory overhead increases. Similarly, when dimension size is 5 the accuracy attained is 0.983 and when it is increased to 11 the accuracy attained is 2.17.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

From this it is clear that accuracy of classification depends on dimension size. The overall result achieved shows scalable performance of ACM model compared with state-of-art model.

Table. 1 Comparison with State of Art Technique for Building Classification Tree

	Random [9]	ANN [9]	ACM-Local	ACM-Hadoop
Total CPU Time (s)	129.69	52.5	35.25	2.37
Average Accuracy	0.977	0.971	0.089	0.989
Memory Overhead (kilo bytes)	0.71	0.69	0.31	0.11
Speedup	14	14	-	16

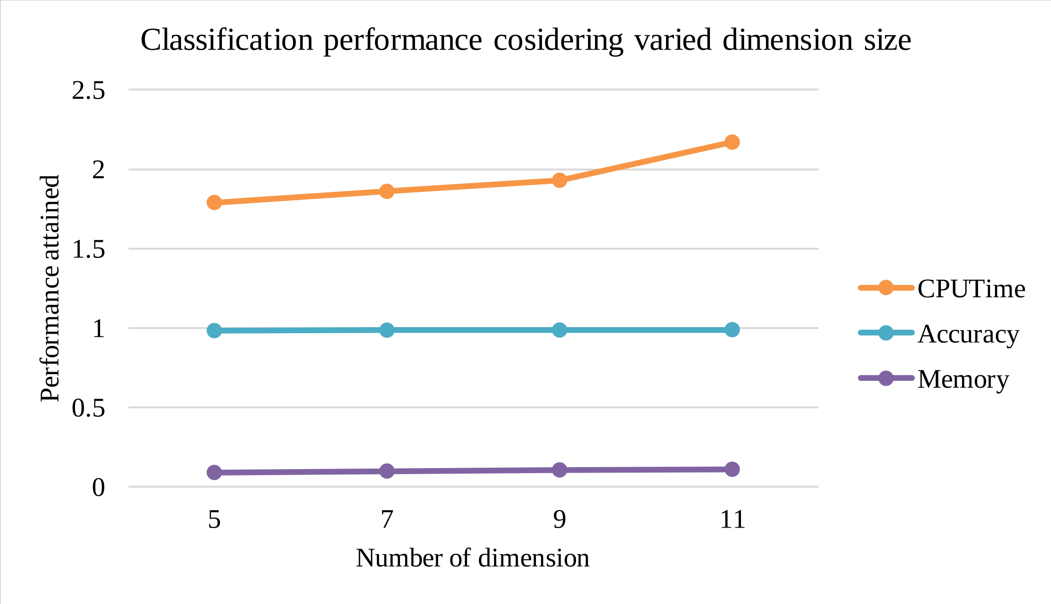

Fig. 6 Classification performance evaluation considering varied dimension size

Table. 2 Classification Performance Evaluation Considering Varied Dimension Size

Dimension size	Total CPU Time (s)	Average Accuracy	Memory Overhead (kilo bytes)
5	1.79	0.983	0.09
7	1.86	0.986	0.099
9	1.93	0.987	0.106
11	2.17	0.989	0.11
Average	1.9375	0.986	0.101

IV. CONCLUSION

This work presented an efficient and accurate classification model for performing analysis on agro related unstructured data. This work presented a selective classification model that perform analysis on multi-dimensional (high dimensional) data. For performing analysis distributed computing multi-level cloud computing framework is adopted. Minimizing cost of processing is most desired on such platform. Therefore, this work presented a hierarchical clustering model to classify data based on user requirement defined. To provide scalable performance for analysis huge high dimensional data parallel clustering algorithm using Hadoop framework is presented. Experiment are conducted on real-time data intensive and scientific application. The outcome shows ACM reduces average computation cost by 35.88%, and latency by 18.39%. Further, ACM-local reduces total CPU time and Memory overhead by 32.85% and 55.07% respectively and improves accuracy by 1.82%. Similarly, ACM-Hadoop classification model reduce total CPU time and Memory overhead by 95.86% and 84.05% respectively, improves accuracy by 1.82% and attain speedup of 16. The overall result achieved shows scalable performance of ACM model compared with state-of-art model in terms of computation cost, latency, total CPU time, accuracy, memory efficiency and speedup. The future work we would consider designing a multi-level cloud storage architecture for cost efficient agriculture management system. Further, we will also considers obtaining some real-time agriculture data or generating agriculture data manually.

REFERENCES

1. Bernard A, “A DSS is an Integration of Web-Based Programs. Geographic Information Systems (GIS) Capabilities and Databases”, USA, pp: 484-495, 2003.
2. Wu J, Ping L, Ge X, et al. Cloud storage as the infrastructure of cloud computing[C]//Intelligent Computing and Cognitive Informatics (ICICCI), 2010 International Conference on. IEEE, 380-383, 2010.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

3. David H, “Web-Based Systems, Client can use any internet-Connected computer or web-enabled Mobile Phone or PDA to gain Real time Access to the data, USA”, 2006.
4. L. Kuang, L. T. Yang, J. Chen, F. Hao and C. Luo, "A Holistic Approach for Distributed Dimensionality Reduction of Big Data," in IEEE Transactions on Cloud Computing, vol. 6, no. 2, pp. 506-518, 2018.
5. Y. Gong, S. Lazebnik, A. Gordo, and F. Perronin, “Iterative quantization: A procrustean approach to learning binary codes for large-scale image retrieval,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 35, no. 12, pp. 2916–2929, 2013.
6. T. Ge, K. He, Q. Ke, and J. Sun, “Optimized product quantization,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 36, no. 4, pp. 744–755, april 2014.
7. L. Bao, Q. Yang, and H. Jin, “Fast edge-preserving patchmatch for large displacement optical flow,” IEEE Transactions on Image Processing, vol. 23, no. 12, pp. 4996–5006, 2014.
8. D.Cozzolino, G.Poggi, and L.Verdoliva, “Efficient dense-field copymove forgery detection,” IEEE Transactions on Information Forensics and Security, vol. 10, no. 11, pp. 2284–2297, 2015.
9. L. Verdoliva, D. Cozzolino and G. Poggi, "A Reliable Order-Statistics-Based Approximate Nearest Neighbor Search Algorithm," in IEEE Transactions on Image Processing, vol. 26, no. 1, pp. 237-250, 2017.
10. Sengupta S, Annervaz K M, Saxena A, et al. Data Vaporizer - Towards a Configurable Enterprise Data Storage Framework in Public Cloud[C]. IEEE International Conference on Cloud Computing. IEEE, 73-80, 2015.
11. X. Ren, P. London, J. Ziani and A. Wierman, "Datum: Managing Data Purchasing and Data Placement in a Geo-Distributed Data Market," in IEEE/ACM Transactions on Networking, vol. 26, no. 2, pp. 893-905, April 2018.
12. “Apache Hadoop.” [Online]. Available: http://hadoop.apache.org/. [Accessed: 21-Oct-2017].
13. Lipeng Wan, Qing Cao, Feiyi Wang, Sarp Oral ”Optimizing checkpoint data placement with guaranteed burst buffer endurance in large-scale hierarchical storage systems,” Journal of Parallel and Distributed Computing, Volume 100, Pages 16-29, 2017.
14. Bharathi S, Chervenak A, Deelman E, Mehta G, Su MH, Vahi K. Characterization of scientific workflows. In: Workflows in Support of Large-Scale Science, 2008. WORKS 2008. Third Workshop on; p. 1±10, 2008.
15. Vinothina.V, R.Sridaran. ”Scheduling Scientific Workflow Based Application Using ACO in Public Cloud” International Journal of Engineering and Technology (IJET), Vol 7 No 6, Dec 2015-Jan 2016.
16. Aggo senosry data. “http://archive.ics.uci.edu/ml/datasets/gas+sensors+for+home+activity +monitoring”, last accessed july 26, 2018.