Journal of Medical Systems (2019) 43: 239
https://doi.org/10.1007/s10916-019-1369-3

TRANSACTIONAL PROCESSING SYSTEMS

Reduced Time Compression in Big Data Using MapReduce Approach and Hadoop

K. Meena [1] · J. Sujatha [2]

Received: 11 March 2019 / Accepted: 5 June 2019 / Published online: 19 June 2019

Abstract
An exponential rise has been observed in the data volume over the time when considering a real time environment. A phenomenal feature termed as 'Predictability' helps in predicting and portraying related data to the user according to their needs. Moreover, classification of Big Data is usually a tedious and lengthy task. The technique of MapReduce Framework performs the data processing that being paralleled by data distribution in small chunks through the clusters. This Map Reduce technique is being proposed which is employed to process heterogeneous data items. Few issues that are being targeted in the existing paper include associating climatologically and meteorological information with large variety of farming decisions. Using the well-known MapReduce framework the above issues and challenges can be resolved. The existing paper proposes empirical techniques of climate classification and prediction by adopting Co-EANFS (Co-Effective and Adaptive Neuro-Fuzzy System) approach for data handling. Furthermore, the paper examines association rule mining too, which is being implemented for examining the best crop production by relying upon the soil and weather condition. Lastly, a technique is proposed for managing various levels such as preprocessing, clustering, classification and prediction. First, the weather dataset is being collected which undergoes processing; thereafter the proposed model is implemented which results in formation of cluster data sets linked to each season. For evaluating the performance, accuracy predictions generated by Co-EANFS is used which being formulated with varying no: of inputs and variables. The proposed framework acquires least execution time.

Keywords MapReduce · Big data · Hadoop · Clustering · Classification · Prediction · Co-effective and adaptive neuro-fuzzy system (Co-EANFS) · Association rule mining

Introduction

Worldwide there exist plentiful volume of data and maintaining it becomes all the more necessary. Hadoop, an Apache framework is adopted for processing, storing and managing bulkier data. The data storage takes place within a distributed environment and so Hadoop contains Hadoop DFS (Hadoop Distributed File System). Existence of this abundant volume of data in a complicated structure is referred to as 'Big Data'. Utilizing the traditional computing methodologies, execution of this big data is not that easy. Hence, there occurs a need of adopting a better technique, the proposed Map reduce framework is such a type. Map reduce is specifically designed to handle and process bulkier data. The processing takes place by dividing the data into different tasks that acts as an input for map tasks and reducer processes, the result is the data received from the mapper. The bulkier data is basically distributed amidst the nodes [1]. In case of any of the node failure within the network, the system still functions correctly. This lowers the risk of catastrophic system failure. Apache Hadoop includes the following: Hadoop kernel, HDFS (Hadoop distributed file system) and map reduce paradigm [2].

Today, there exists plentiful volume of data. Computing and managing it becomes all the more necessary as it involves immense processing time. Hadoop, an Apache framework is a solution for processing, storing and managing bulkier data.

This article is part of the Topical Collection on *Transactional Processing Systems*

✉ K. Meena
meen.nandhu@gmail.com

J. Sujatha
jsujathacse@gmail.com

[1] Department of Computer Science & Engineering, Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Chennai, India

[2] Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Chennai, India

Few issues that are being targeted in the existing paper includes associating climatologically and meteorological information with large variety of farming decisions. The existing paper proposes empirical techniques of climate classification and prediction by adopting Co-EANFS (Co-effective and adaptive neuro-fuzzy system) approach for data handling. Furthermore, the paper examines association rule mining too, which is being implemented for examining the best crop production by relying upon the soil and weather condition. Finally this paper proposes a mechanism for handling stages like preprocessing, clustering, classification and prediction. First, the weather dataset is being collected which undergoes processing, thereafter the proposed model is implemented which results in formation of cluster data sets linked to each season. The initial step of preprocessing involves cleaning of original data, thereafter clustering of data takes place in which cluster data are formed linked to each season. Co-EANFS (a ML technique) is employed for classification. The approach of Co-EANFS helps in analyzing data as well as comprehends it for drawing future predictions, and is also beneficial in weather forecasting. The blend of neural networks and forms the Co-EANFS which results in a hybrid system capable of comprehending automatically and adapting data. Co-EANFS is beneficial in predicting about the weather forecast. Furthermore, the paper adopts association rule mining which being a technique of data mining and helps in identifying frequent patterns from data sets prevailing in various databases. It takes into account the attributes and their relationship for examining the best crop production by relying upon the soil and weather condition. According to the output the Co-EANFS yields in high accuracy and consumes low time compared to rest approaches.

Organization of the paper

The paper is structured as follows. Second section illustrates the literature survey. Third section lays down the proposed HFDS and predicting future forecasting relying upon previous data and aspects of different stages. Fourth section, illustrates experimental outcome. Fifth section brings about the conclusion along with suggesting future research work.

Systematic literature survey

Saraladevi et al. [1] presented different challenges related to big data. The terminology of big data deals with storage of abundant quantity of data gathered from different kind of sources. Following are the challenges concerning big data that's being presented in the paper, namely - storage, security, processing and management issues. In order to safeguard the data in HDFS, a list of techniques is being adopted by the author that relies upon Kerberos, Bull eye algorithm and name node. Kerberos permits only authenticated users to access Hadoop DFS. The Bull eye algorithm describes the application of security from one node to the other. The above mentioned algorithms are employed in the base layer of HADOOP.

Anam Alam, 2014 [2] represented the vision of HADOOP architecture along with its uses. The application of HADOOP is in a wide domain such as social media, data intensive applications, data analytics etc. In social media it's being imbibed for opinion mining. The drawback confronted by Hadoop is of incremental computations, which the author has resolved using caching [3]. HADOOP DISTRIBUTED FILE SYSTEM: HADOOP is divided into 2 parts: first, HDFS (Hadoop Distributed File System) and the second is Map reduce. HDFS basically performs storage of data whereas map reduce performs the data processing. Nusrat Sharmin Islam et al. [3] proposed a set of data access techniques for effectively accessing the Hadoop DFS. The heterogeneity of data is taken into account. The author claims that using the proposed data access strategies, the access performance has increased to around 33% compared to the existing locality aware data access. Also the study reveals that execution time is reduced by 32% and Spark sort by 17%.

MAPREDUCE: The Map reduce component of Hadoop is responsible for processing bulky data. Hirotaka Ogawa et al. [4] proved that the shortcoming of current map reduce framework in the paper and proposed a novel map reduce framework referred to as SSS which relies upon distributed key- value store. Here, the word count and iterative composite forms the two benchmarks as presented by the author. The outcome reveals that the proposed SSS framework is 1- 10 times faster compared to the traditional HADOOP. SCHEDULING IN HADOOP: Kamal Kc et al. [5] represented the deadline constraint scheduler in the present work. Hadoop makes use of FIFO as the conventional scheduler. In the new approach the author has built a scheduler that considers the user mentioned deadlines for scheduling any job. It's made sure to schedule only those specific jobs whose deadline can be fulfilled. In case of different deadlines for different jobs the scheduler allocates a set of tasks to the task tracker, as a result it can be ensured whether the deadline is fulfilled or not [6].

Soumadip Ghosh et al. [7] proposed the NF (neuro-fuzzy) classification approach which identifies different soil classes from huge imagery soil databases. Feature-wise degree of belongings of the imagery databases are considered for acquiring soil classes with the help of fuzzification method. The technique is being implemented on three UCI databases, these are: Forest Cover type, Stat log and Lands at Satellite along with Wilt for the purpose of soil classification technique. Their performance is then compared with four popular classification

algorithms. The above mentioned measure confirms the superiority of neuro-fuzzy approach [8].

Geoff Kuehn et.al presented [8] ADOPT (Adoption and Diffusion Outcome Prediction Tool). Though there have been an intense study and acquiring of facts and parameters that impacts adoption of new strategies in the field of agriculture but hardly any effort is put to build predictive quantitative-models-of-adoption by the ones performing agricultural research, development, extension and policy. ADOPT is basically framed to improvise the conceptual comprehension and consideration of the adoption process by the individuals responsible for agricultural research, development, extension and policy.

Modise Wiston et.al proposed NWP (Numerical weather prediction) a contemporary forecasting technique formulated via simplified systems of physical laws of the atmosphere. In olden times, people relied upon their instinct and experience for predicting the weather and to conclude when and what is going to happen. But with changing times and surroundings it was clear that, predicting accurate weather changes can't be done merely on the basis of knowledge and experience. Here the approach of NWP is proposed for providing a solution for the weather prediction [9].

Shah Dhairya Vipulkumar et.al., discussed few novel data placement approaches that can be implemented in heterogeneous cloud scenario for increasing the processing of big data by improving the response time. The work presents different novel techniques/approaches for enhancing data placement in a heterogeneous cloud scenario for increasing the processing of big data [10].

Abhishek et al. [11] proposed Artificial Neural Networks (ANNs) in order to predict monthly average rainfall of a place within India depicted using monsoon type climate. The data being utilized in the study was of 8 months yearly. As during these specified months there is maximum possibility of rainfall occurrence. The descriptive variables being used were humidity and average wind speed. Three different networks were utilized to perform the experiment namely: Feed Forward Back Propagation, Layer Recurrent, and Cascaded Feed Forward Back Propagation. Thereafter, output retrieved from each network was compared, and it was revealed that the Feed Forward Back Propagation network yielded in great results.

R. Samyaet. et.al., surveyed different forecasting techniques related to cloudburst with the help of Data Mining and ANN. Weather Forecasting tends to be a technical and scientific concern over the years. Among various weather related issues, Cloudburst being a matter of prime concern as it leads to maximum destruction. Researchers opt for NWP (Numerical Weather Prediction) method for prediction of such challenging weather conditions. The work reveals that the DM approaches such as Fuzzy logic, ANN and ANFIS yields in high accuracy [12].

Ashwani Kumar et al. [13] have proposed a new "Agro algorithm" that's being employed for processing of massive agricultural data by adopting the Big data Hadoop platform. Using this technique in agricultural domain, the crop production can be predicted, also best crop can be suggested to the farmer for increasing the profit and quality. Also, which soil is apt for which specific crop can be analyzed. Climatic condition holds very significant when considering farming hence which crop is suitable for which soil needs to be predicted in order to enhance the quality of crop by incorporating weather and disease based data sets.

Ankalaki et al. [14] put forth a comparative study on DBSCAN and AGNES techniques for the process of clustering. Multiple Linear Regression (MLR) is adopted for predicting Crop production along with deriving a formula for every crop. S. Poongodi et al. [15] proposes that depending on this data, the farmer can select best suitable crop for the process of cultivation. Feature selection employs genetic algorithm in order to choose optimal features. Thereafter, enhanced C4.5 along with ANFIS classifier is adopted for data classification on region wise basis. The proposed work yields an increased accuracy of 92.50% compared to already available classifier.

D. Ramesh et al. [16] presented DM techniques such as classification and clustering and NN (Neural Networks) to illustrate DM applications in the realm of Agricultural Yield. The proposed K-Means Algorithm and MLR (Multiple Linear Regression) makes use of rainfall as the dependent attribute whereas the independent attributes being - Year, Area of sowing and Production for drawing out their prediction. It was revealed that the MLR approach yielded 98% of accuracy which was higher compared to K-Means Algorithm where the accuracy achieved was 96%.Farming represents the backbone of Indian economy. Crop production primarily relies on weather parameters like change in climate, rainfall, soil etc... that mainly affects and enhances the harvest. For the process of forecasting, the prevailing technique for crop yield prediction adopts DM techniques. Existing paper work illustrates few current techniques that are being adopted for predicting crop yield [17].

S. Dahikar and S. Rode proposes crop prediction strategies which can be incorporated for detecting reasonable crop yield by identifying various soil related and other factors prevailing in the environment. And for the same purpose the ANN approach is being adopted. The capability of ANN is being employed to analyze and anticipate the cop yield for any region in the country. It was revealed that ANN tends to be beneficial for the forecasting of crops. The paper work utilizes the factors concerning the provincial soil factor. Forward back propagation approach is employed for analyzing the same. For

achieving competency, Mat lab Artificial Neural Network is used [18].

Proposed work

Overview

Existence of this abundant volume of data in a complicated structure is referred to as 'Big Data'. In the field of Agriculture, big data is an assembly of structured and unstructured form of data. The Map Reduce framework of HADOOP aids in processing of bulkier data effectively. It splits the data processing task in various stages referred to as set of Maps and Reduces which are being worked upon simultaneously/ parallel on multiple systems. Collection of these agriculture datasets is performed online via web portal. Co-EANFS is considered as the most effective crop detection and classification technique that incorporates ML techniques. It's basically based on layers so that the appropriate crop can be categorized from actual datasets. Further Association rule mining is being incorporated for examining the best crop production by relying upon the soil and weather condition. Lastly, the work focuses to compute performance metric in overall assessment outcome. Figure 1 shows the overall proposed architecture of the system.

Collection of datasets

Collection of these heterogeneous datasets is performed online via UCI website. The weather data gathered is of over last 3 years and Co-EANFS is adopted for weather prediction and identification and for suggesting the best suitable crop on the basis of soil and Weather circumstances. The paper work resolves the issue of anticipating accumulated Weather precipitation for the subsequent day, relying upon the data that's collected in prior days. Various meteorological parameters of a given place and time build up a weather, these conditions may be temperature, wind, sunshine, rain, snow etc. at a particular time and place. On the other hand 'climate' symbolizes a relatively long duration traits of a weather being noticed of a particular place. Climate depends on the whole ecosystems, livings, farming, and communities of an area. Hence it can be stated that the

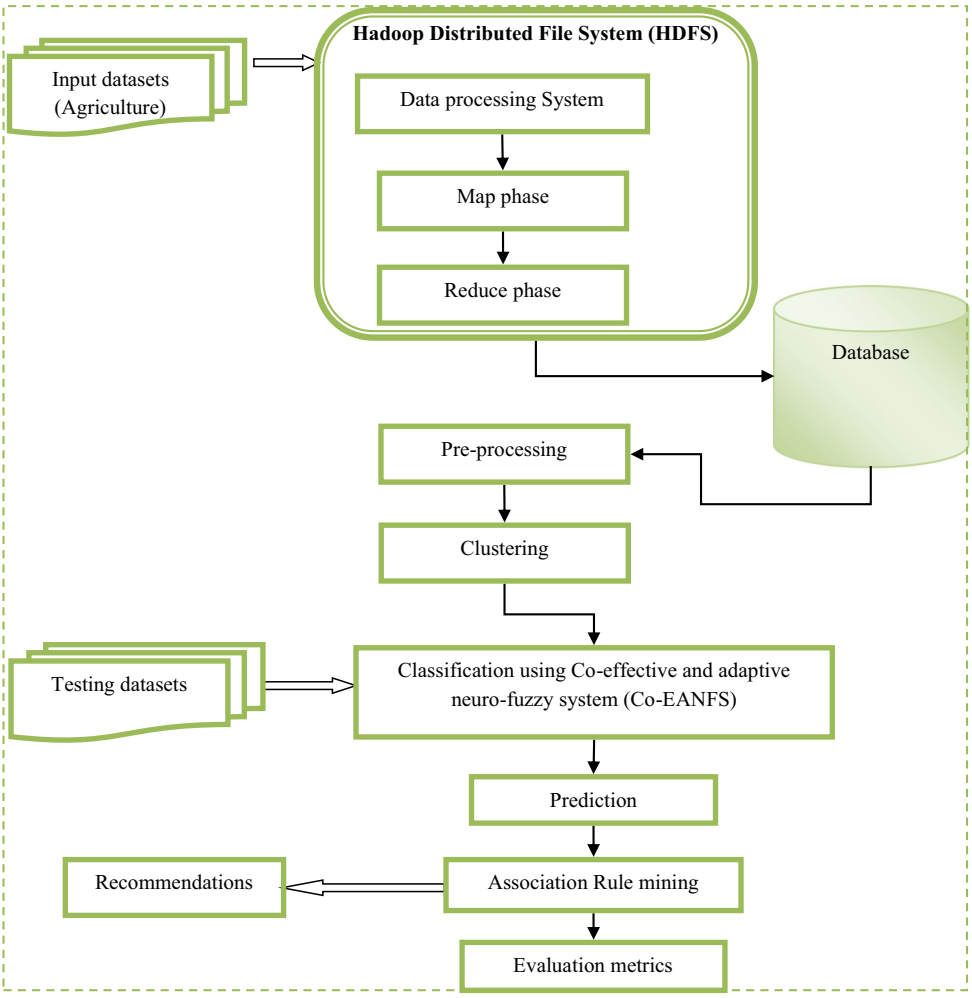


Fig. 1 Overall proposed architecture

climate is an accumulation of long-term weather conditions drawing out the average conditions and their variability.

Figure 2a, b, c depicts weather-crop condition data, weather condition and soil datasets respectively. Weather and soil condition datasets are gathered anticipating the accumulated Weather precipitation for subsequent day, relying upon the data that's collected in prior days. Various meteorological parameters of a given place and time build up a weather, these conditions may be temperature, wind, sunshine, rain, snow etc. at a particular time and place. In the realm of agriculture, Crop prediction is considered as a major theme that's being utilized for yield estimation, yield mapping, matching supply of crop supply with that of the demand and crop management for improvising productivity. The aim of Weather Prediction is to anticipate the atmospheric condition concerning the future and place/area. Various enterprises and agricultural domains intensely relies upon weather conditions. Basically the Weather Predictions helps in preparing for future upcoming natural calamities that is usually a result of some fluctuations in climate.

Hadoop distributed file system (HDFS)

To for the management and handling of massive volume of data the approach of load balancing is being utilized, where data load on single server is minimized by distributing data or spreading it out across multiple servers. The technique of HADOOP helps in managing this 'big data'. Hadoop, an Apache framework is adopted for processing, storing and managing bulkier data. It involves voluminous nodes and data in terabytes. The bulkier data is basically distributed amidst the nodes. In case of any of the node failure within the network, the system still functions correctly. This lowers the risk of catastrophic system failure. Apache Hadoop includes the following: Hadoop kernel, HDFS (Hadoop distributed file system) and map reduce paradigm. The HDFS (Hadoop Distributed File System) is a fault-tolerant storage system which is capable of storing abundant quantity of information that can be raised with no system failure of loss of data.

- **HDFS contains three major components:**

– Name node
– Data Node
– Secondary Name node

- **Map stage:** Mapper / map is responsible for processing the input data which is basically a file/directory that is being saved in HDFS. The input data (i.e., the file) is forwarded line by line to the to the mapper function. Once the mapper performs data processing, it generates little data chunks.
- **Reduce stage:** Reduce stage is a blend of the Shuffle and the Reduce stage. It fetches the data from the mapper and processes it further and generates a new result which is being stored in Hadoop DFS.

Map reduces Algorithm

```
Input: // Weather dataset
Begin
1: class Mapper
2: method Map(LongWritable, Text, Text, IntWritable)
3: for all Text, IntWritable do
4: Emit(string temp; line 10)
5: if stringUtils is Numeric then
6: output (datepart , Temperture)
End.
Begin
7: class Reducer
8: method Reduce (Text, IntWritable, Text, IntWritable)
9: sum 0
10: while all count temp counts [temp1; temp2; and tempn] do
11: sum = sum + temp
12: Emit (temp ; count sum)
End
```

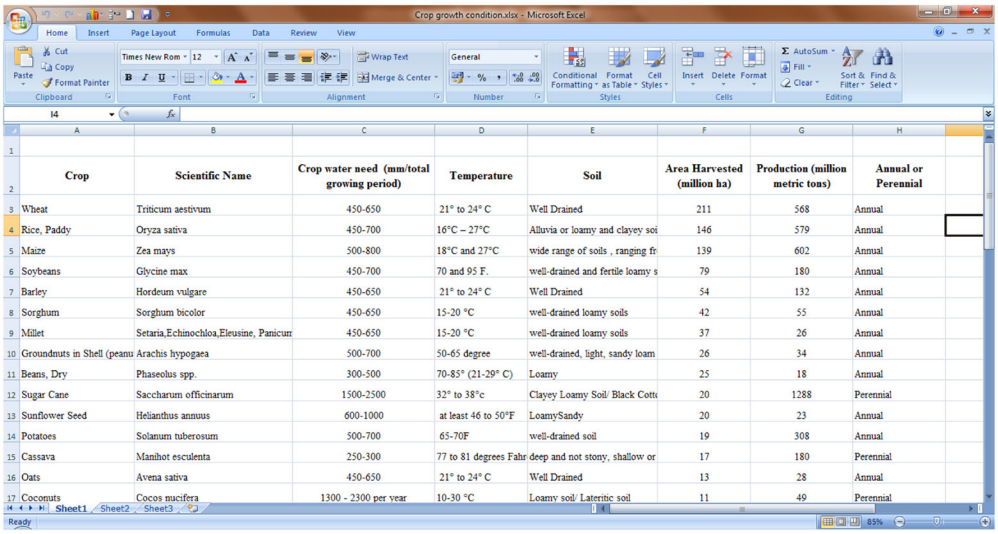
Preprocessing

Data pre-processing helps in enhancing the input quality that further impacts the analytical functionality and efficiency. Weather datasets that are being assembled undergo pre-processing, that is it eliminates unnecessary data from weather datasets that is either duplicate, involves same weather types or is not consistent. The process of Data preparation is compulsory. Here previous unwanted data is transformed into new data that is suitable for DM (Data Mining) task. In case the data is not prepared, it's not taken up by the DM algorithm for processing or if taken, reports error when it's executed. The DM algorithm might work in the best case, but the output generated may be crap or inaccurate.

Preprocessing involves the following steps:

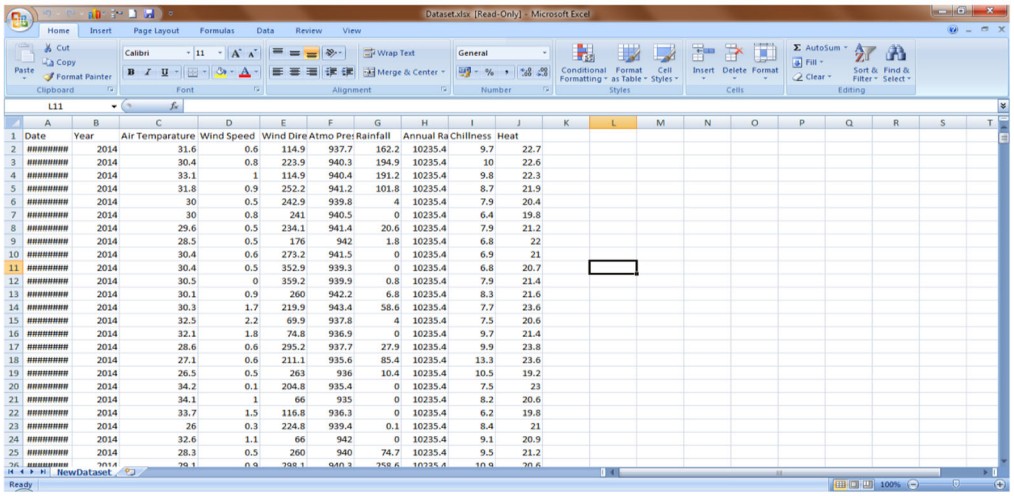
1. Eliminating unnecessary data and preserving the significant one for eg: membership ID and the member's name.
2. Performing conversion of data formats to improve the attributes, for instance converting sales time into hours, denoting the period.
3. Detecting and examining that information which doesn't fulfill further analysis.
4. Removing information that is undistinguished.

At this level, consistent data model is formulated that handles duplicate and missing data and distinguishing junk data.



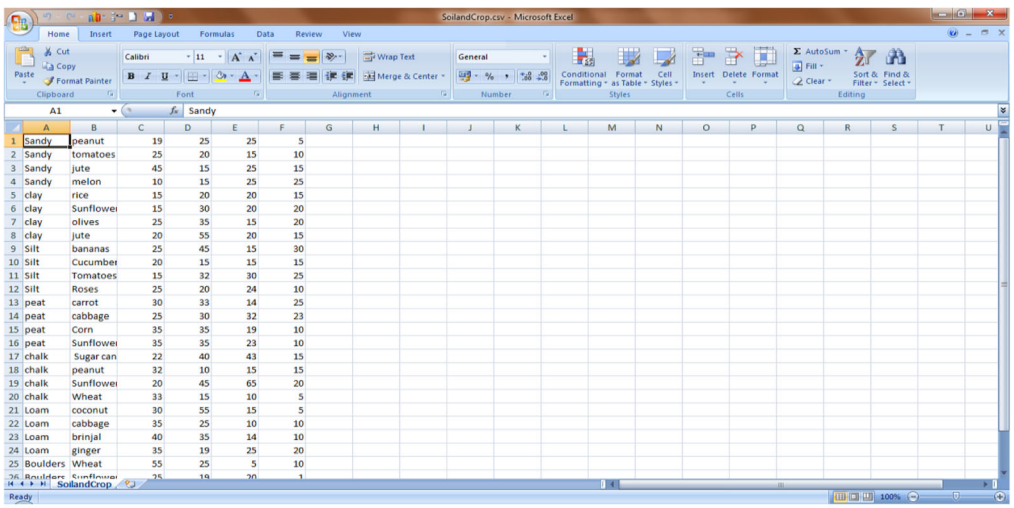
Crop	Scientific Name	Crop water need (mm/total growing period)	Temperature	Soil	Area Harvested (million ha)	Production (million metric tons)	Annual or Perennial
Wheat	Triticum aestivum	450-650	21° to 24° C	Well Drained	211	568	Annual
Rice, Paddy	Oryza sativa	450-700	16°C – 27°C	Alluvia or loamy and clayey soi	146	579	Annual
Maize	Zea mays	500-800	18°C and 27°C	wide range of soils , ranging fr	139	602	Annual
Soybeans	Glycine max	450-700	70 and 95 F.	well-drained and fertile loamy s	79	180	Annual
Barley	Hordeum vulgare	450-650	21° to 24° C	Well Drained	54	132	Annual
Sorghum	Sorghum bicolor	450-650	15-20 °C	well-drained loamy soils	42	55	Annual
Millet	Setaria,Echinochloa,Eleusine, Panicum	450-650	15-20 °C	well-drained loamy soils	37	26	Annual
Groundnuts in Shell (peanu	Arachis hypogaea	500-700	50-65 degree	well-drained, light, sandy loam	26	34	Annual
Beans, Dry	Phaseolus spp.	300-500	70-85° (21-29° C)	Loamy	25	18	Annual
Sugar Cane	Saccharum officinarum	1500-2500	32° to 38°c	Clayey Loamy Soil/ Black Cott	20	1288	Perennial
Sunflower Seed	Helianthus annuus	600-1000	at least 46 to 50°F	LoamySandy	20	23	Annual
Potatoes	Solanum tuberosum	500-700	65-70F	well-drained soil	19	308	Annual
Cassava	Manihot esculenta	250-300	77 to 81 degrees Fahr	deep and not stony, shallow or	17	180	Perennial
Oats	Avena sativa	450-650	21° to 24° C	Well Drained	13	28	Annual
Coconuts	Cocos nucifera	1300 - 2300 per year	10-30 °C	Loamy soil/ Lateritic soil	11	49	Perennial

(a)



Date	Year	Air Temparature	Wind Speed	Wind Dire	Atmo Pres	Rainfall	Annual Ra	Chillness	Heat
#########	2014	31.6	0.6	114.9	937.7	162.2	10235.4	9.7	22.7
#########	2014	30.4	0.8	223.9	940.3	194.9	10235.4	10	22.6
#########	2014	33.1	1	114.9	940.4	191.2	10235.4	9.8	22.3
#########	2014	31.8	0.9	252.2	941.2	101.8	10235.4	8.7	21.9
#########	2014	30	0.5	242.9	939.8	4	10235.4	7.9	20.4
#########	2014	30	0.8	241	940.5	0	10235.4	6.4	19.8
#########	2014	29.6	0.5	234.1	941.4	20.6	10235.4	7.9	21.2
#########	2014	28.5	0.5	176	942	1.8	10235.4	6.8	22
#########	2014	30.4	0.6	273.2	941.5	0	10235.4	6.9	21
#########	2014	30.4	0.5	352.9	939.3	0	10235.4	6.8	20.7
#########	2014	30.5	0	359.2	939.9	0.8	10235.4	7.9	21.4
#########	2014	30.1	0.9	260	942.2	6.8	10235.4	8.3	21.6
#########	2014	30.3	1.7	219.9	943.4	58.6	10235.4	7.7	23.6
#########	2014	32.5	2.2	69.9	937.8	4	10235.4	7.5	20.6
#########	2014	32.1	1.8	74.8	936.9	0	10235.4	9.7	21.4
#########	2014	28.6	0.6	295.2	937.7	27.9	10235.4	9.9	23.8
#########	2014	27.1	0.6	211.1	935.6	85.4	10235.4	13.3	23.6
#########	2014	26.5	0.5	263	936	10.4	10235.4	10.5	19.2
#########	2014	34.2	0.1	204.8	935.4	0	10235.4	7.5	23
#########	2014	34.1	1	66	935	0	10235.4	8.2	20.6
#########	2014	33.7	1.5	116.8	936.3	0	10235.4	6.2	19.8
#########	2014	26	0.3	224.8	939.4	0.1	10235.4	8.4	21
#########	2014	32.6	1.1	66	942	0	10235.4	9.1	20.9
#########	2014	28.3	0.5	260	940	74.7	10235.4	9.5	21.2

(b)



Sandy	peanut	19	25	25	5
Sandy	tomatoes	25	20	15	10
Sandy	jute	45	15	25	15
Sandy	melon	10	15	25	25
clay	rice	15	20	20	15
clay	Sunflowe	15	30	20	20
clay	olives	25	35	15	20
clay	jute	20	55	20	15
Silt	bananas	25	45	15	30
Silt	Cucumbe	20	15	15	15
Silt	Tomatoes	15	32	30	25
Silt	Roses	25	20	24	10
peat	carrot	30	33	14	25
peat	cabbage	25	30	32	23
peat	Corn	35	35	19	10
peat	Sunflowe	35	35	23	10
chalk	Sugar can	22	40	43	15
chalk	peanut	32	10	15	15
chalk	Sunflowe	20	45	65	20
chalk	Wheat	33	15	10	5
Loam	coconut	30	55	15	5
Loam	cabbage	35	25	10	10
Loam	brinjal	40	35	14	10
Loam	ginger	35	19	25	20
Boulders	Wheat	55	25	5	10

(c)

Fig. 2 **a**, **b**, **c** Weather crop condition datasets

At last the cleaned data is converted to a format that being appropriate for the process of data retrieval.

Clustering

Clustering process clusters a group of features in a manner that features belonging to the same cluster are majorly equivalent (to an extent) to one another compared to the features belonging to other clusters. It's an undertaking of exploratory DM (data mining) and common methodology for statistical DM, that being adopted in many endeavors. Clustering is implemented in meteorological application for deciding upon grouping of weather conditions. It takes into account parameters like humidity, air temperature, rainfall, sunlight etc. here clustering is done relying upon seasons based clustered that involves summer, rainy and winter season. It holds clusters that are related/similar thereby storing one segment which are being clustered for precise results classification (Figs. 3 and 4).

Classification

Classification is an ordered set of similar classes that performs grouping of data based on their similarities. There are codes and descriptors in it and assigns survey responses into significant categories for generating meaningful data. The classification in agriculture weather classification relies upon the condition of weather and soil. Training and testing are the two sections of classification. The paper work proposes C0-EANFS for classification of weather condition.

Co-effective and adaptive neuro-fuzzy system (Co-EANFS)

Co-EANFS fits in the general class of ANFS (adaptive neuro-fuzzy systems). C0-EANFS is considered to be a universal approximate of any non-linear function. Co-EANFS characteristics are highlighted using the benefits of Integrating NN (neural network) along with FS -fuzzy systems referring the same topology. Here the author proposes C0-EANFS which is an extension of Neuro-fuzzy based technique, for classification of weather condition. Co-EANFS's is capable enough as it handles pattern-depended weights amidst the consequent layer and the fuzzy association layer. A membership value signifies to the dynamic changeable weights based on input patterns. Whereas, the back propagation method having sigmoid neuron functions universally updates weight co-efficient related to each input pattern, trying to identify one particular set of weights that's common in all training patterns. NN (neural networks) when locally adjusted with normalization results in extrapolation outcome that's analogous to the back propagation.

Algorithm of Co-EANFS:

```
Function (Attributes, btraining, threshold, DL _ attributes);
Benign;
Inputs: No of Classes (Attributes);
 Threshold;
 Training Data;
Sort training data according to class
If  Length (Training set < Threshold)
      Generate supporting data and add to training set
      Train Database with training set;
      Decision List = MakeEmptyDL () do
      Decision List = Search (MakeEmptyRule (), database);
      Sort Decision List
      Add default rule to dicisionList
End
All classes (attributes) to convert == Kernel Function;
Kernel_Function(); // Activation
If (attributes==kernel values<=length (Threshold (i))
 Trained successfully;
Passing to testing process;
End
If (Training &&Testing<=Length (threshold (i))) && (Training
&&Testing==Length
(threshold (i))) then
      Similar attributes based DicisionList of Result true;
      Else
      Dissimilar attributes based DicisionList of Result false;
End
```

Step 1: Every node-i in present layer is adaptive having a node function. Nodes generate Membership values.

$$M_i^1 = \mu_{P_i}(A) \text{ And } M_i^2 = \mu_{Q_i}(B) \tag{1}$$

Here, x denotes input to node-i, p_i denotes linguistic variable linked with the node function and μP_i denotes membership function of P_i,

μ_{P_i} ➔ Input parameters, attributes like,

Just for illustration purpose, membership functions are being denoted by the bell-shaped function given by:

$$\mu_{P_i}(A) = \frac{1}{1 + \left|\frac{x - r_i}{s_i}\right|^{2t_i}} \tag{2}$$

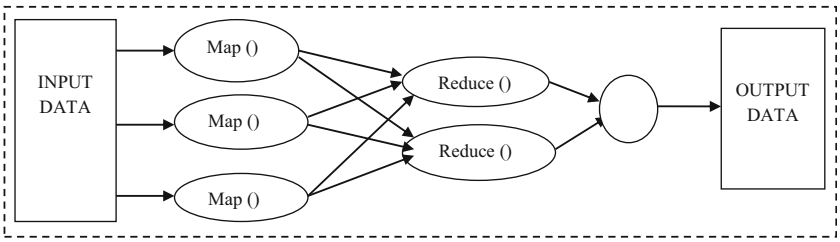


Fig. 3 Hadoop server

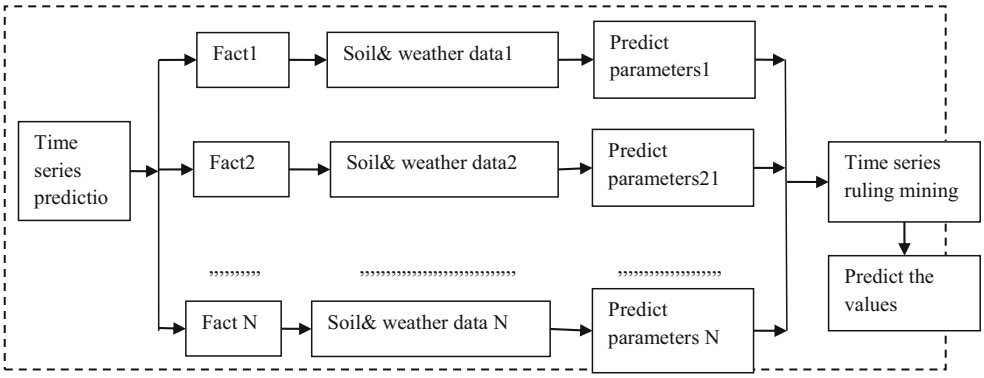


Fig. 4 Association rule mining to prediction values

Here, s_i, t_i, r_i signifies parameters to be learnt and are premise parameters.

Step 2: In this layer, every node is a fixed node which computes the firing strength-w_i of a rule. Result of every node is the product of entire data coming to it and is denoted as,

$$w_i = \mu_{P_i}(A) * \mu_{q_i}(B) \qquad i = 1, 2, , \tag{3}$$

Step 3: In this layer, every node is a fixed node. Every i^{th} node computes the ratio of the i^{th} rule's firing strength to the total of entire rules firing strengths. Result from the i^{th} node is the normalized firing strength denoted by,

$$\overline{w}_i = \frac{w_i}{w_1 + w_2}, \quad i = 1, 2.. \tag{4}$$

Step 4: Every node- i in present layer is adaptive having a node function denoted by,

$$\overline{w}_i f_i = \overline{w}_i (s_i A + t_i B + r_i), \quad i = 1, 2.. \tag{5}$$

here $\overline{w}_i$ denotes Layer 3 result and $\{s_i, t_i, r_i\}$ represents consequential set of parameter.

Step 5: Single summation node Layer: There is only a single fixed node in this layer which computes the entire output as the total of all incoming data's, i.e.,

$$M_i^1 = Overall\ Output = \sum_i \overline{w}_i f_i = \frac{\sum_i w_i f_i}{\sum_i w_i} \tag{6}$$

Identification of consequent parameters is performed using least squares estimate, concerning the forward pass of the learning algorithm. The backward pass propagates backward the error signals (that being derivatives of the squared error in terms of each node output), from the output layer to the input layer.

Prediction using association rule mining

Using the computer technology, weather prediction can be performed. The Forecast models (complicated computer programs) that execute on super computers are being adopted offers predictions on various atmospheric variables like pressure, temperature, wind pressure, rainfall etc... With evolution and growth in technology, abundant quantity of data is being generated which needs to be effectively stored for future analysis and pattern identification within the data. As a result a 'big data' framework is being constructed that is highly efficient in storage of massive amount of data. Association rule mining can then be implemented on such data for carrying out prediction. It's revealed that association rule mining generates high accuracy. The intelligent agriculture system tends to employ the latest technique to enhance the Weather prediction, making sure that requirements and growth of smarter cities has been made better.

Time series

The performance of prediction model enhances with features that improvises its adapting potential. The study takes into account the features which being the statistical indicators computed for multiple time frames and discussed as below:

Data is assembled from past 3 years of UCI dataset with the help of Association rule mining for predicting weather and suggesting best suitable crop on the basis of Weather conditions.

$$\text{Year of Maximum Weather conditions} = \text{Max}\sum_n^{i=0} n_{Max}(t) \tag{7}$$

$$\text{Year of Minimum Weather conditions} = \text{Min}\ \sum_n^{i=0} n_{Min}(t) \tag{8}$$

$$\text{Year of Average Weather conditions} = \text{Average}\sum_n^{i=0} n_{avg}(t) \tag{9}$$

$$\text{Time series} = (\text{x(t}-1) - \text{x(t)}) + \text{x(t)}, \tag{10}$$

this is Time series equation.

$$\begin{aligned} &\text{Feature Predict on time series} = \text{F(x)} \\ &\text{F(x)} = \left(\text{Max}\sum_n^{i=0} n_{\text{Max}}(t) - \text{Average}\sum_n^{i=0} n_{\text{avg}}(t)\right) + \text{Min}\sum_n^{i=0} n_{Min}(t) \\ &\text{F(x)} = \text{Accurate weather condition.} \end{aligned} \tag{11}$$

The time series prediction is examined so as to identify the influence factors concerning the time series, thereafter researching on influence factors of targeted time series. Related factor time series data is then gathered via Internet,

examined and then pre-processing the historical data. Artificial neuron being a perception, neuron output is computed by:

$$y = f\left(\sum_{i=1}^{n} w_i x_i - \theta\right) \quad (12)$$

here, θ denotes bias, y denotes output, and n signifies number of input $(x_1,\cdots,x_2)$ and weights $(w_1,\cdots,w_n)$. f denotes activation function. ARM makes use of maximum utilized activation function - sigmoid, defined by:

$$(\mathrm{x}) = \frac{1}{1+e^{-x}} \quad \mathrm{x}\in(-\infty,\infty) \quad (13)$$

for updating weight values, back propagation utilizes the output and estimated output by making use of gradient descent. Error E is computed by:

$$E = \frac{1}{2\pi}\sum_{i=1}^{T}\sum_{i=1}^{m}\left(p_i^t - v_i^t\right)^2 \quad (14)$$

Where, T represents number of patterns that were used in training, m represents number of outputs, P denotes desired output and v denotes estimated output. The predictive values (related to influence factors) as the time-series of network input values, adopts multiple network for generating predicted value.

Recommendation

Successful recommendations can be possibly made by the system for the users (with missing prior interest data), relying upon user's preference data and demographics by making use of previous recommendation experiences of similar kind of users. The proposed recommendation methodology resolves the data sparsely problem confronted in recommender systems by its inference skills. Co-EANFS regularization is adopted for implementation and association rule mining for prediction to enhance recommender system's success ratio.

Evaluation metric

Worldwide there exist plentiful volume of data, processing and maintaining it becomes all the more necessary. Hadoop, an Apache framework is adopted for processing, storing and managing bulkier data. The existing paper proposes a map reduce framework that takes in account heterogeneous data within agricultural domain of classification by adopting Association rule mining algorithm and Co-EANFS utilizing best crop prediction on the basis of weather and soil condition datasets. For evaluating performance and computing system stability few parameters are examined and computed that are given below:

For the evaluation of the proposed classification technique, RMSE (Root Mean Square Error), Recall, precision, accuracy are being utilized.

Table 1 Comparison of Hadoop files system

Techniques	CPU utilization	Processing time (ms)
Hash algorithm	Max	2.03
Collaborative filtering	Max, min	1.80
Map reduce technique (HFDS)	Low, min	1.10

True Positive (TP) - If the case in point being positive (classified results) result is classified as positive.
False Negative (FN) - If the case in point being positive (classified results) result is classified as negative.
True Negative (TN) - If the case in point being negative (classified results) result is classified as negative.
False Positive (FP) - If the case in point being negative (classified results) result is classified as positive.

A common approach for evaluating a classified is by calculating the deviation of the classified referring to the actual/true value which actually is the base for RMSE.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(Y_i - R_i\right)^2} \quad (15)$$

Performance measures being adopted assess these algorithms, contain their root. Accuracy is a well utilized measure which being the fraction of correct-classified to the total-possible classifications.

Confusion Matrix

Actual/Predicted	**Negative**	**Positive**
Negative	True Positive	True Negative
Positive	False Negative	False Positive

Confusion matrix is adopted for deriving various performance measures. Concerning the recommender system's DM (data mining) task, an algorithm's performance relies upon its proficiency in learning valuable patterns from the data set. Accuracy is a well utilized measure which being the fraction of correct-recommendations to the total-possible recommendations.

Table 2 Comparison of classification techniques

Classification technique	Accuracy (%)	Efficiency (%)	Time (Ms)
SVM	88.4	92.2	0.58
RBFN	75.7	87.3	0.97
K-NN	76.8	88.1	0.70
Co-EANFS	95.1	98.6	0.45

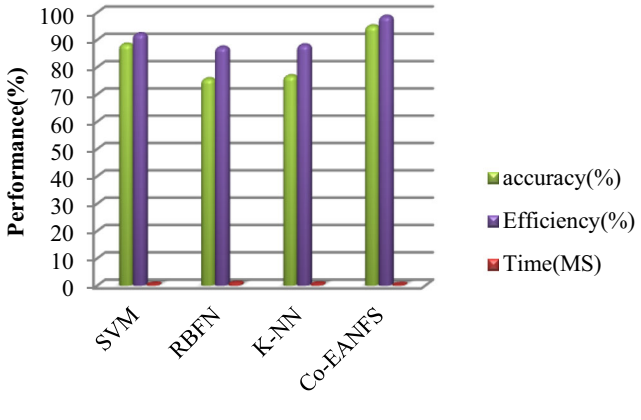


Fig. 5 Comparative techniques of classification process

F-measure is a commonly used single-valued measure which is stated as the harmonic mean of precision and recall.

$$Accuracy = \frac{Correct\ Recommendations}{Totla\ Possible\ recommendations} = \frac{TP + TN}{TP + FP + TN + FN} \quad (16)$$

$$Recall = \frac{Correctly\ recommended\ crop}{Total\ useful\ recommendation} = \frac{TP}{TP + FN} \quad (17)$$

$$Precision = \frac{Correctly\ recommended\ Items}{Total\ recommended\ items} = \frac{TP}{TP + FP} \quad (18)$$

Popular single-valued measure is the F-measure. It is defined as the harmonic mean of precision and recall.

$$F{-}score(efficiency) = 2 * \frac{Recall * Precision}{Recall + Precision} \quad (19)$$

Results and discussions

Datasets holding weather prediction and recommendation based on soil and weather condition are gathered online from the UCI datasets. The paper work is framed and imbibed considering 'Big data' territory, evaluating massive data volume all the more efficiently. Using the following configuration the experiments were carried out: OS - Windows 7, Intel Pentium (R), CPU - G2020 with processer speed 2.90 GHz. Software configuration needed is as follows: OS - Windows 7 and VMware configuration, Front End - JAVA, and HADOOP environment.

Table 3 Comparison of prediction techniques

Prediction techniques	Accuracy (%)
C5	86.1
Numerical weather prediction	88.2
Deep learning prediction	92.1
Association rule mining	93.9

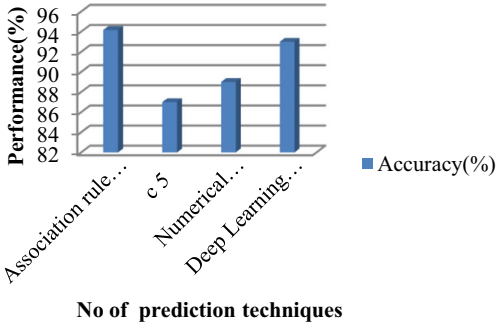


Fig. 6 Prediction techniques performance

Implementation in the research paper work is performed in 'big data', JAVA environment and Hadoop domain. The terminology of big data deals with storage of abundant quantity of data gathered from different kind of sources. In the realm of agriculture, Big Data signifies vast measuring of heterogeneous data, Weather immensely impacts our day to day living and can even bring about everlasting changes to our society, economy and environment at large. Usually the farmers cultivate and harvest their crops relying upon the suggestions and experiences of older generation. They overlook any weather forecast since they trust their traditional approach and knowledge, resultant they confront unexpected loss and damage. Hence, there occurs a need of weather and soil prediction which is a prevailing issue all over. The above issue is resolved in the research work by proposing the Map reduce framework and Co-EANFS (Co-Effective and Adaptive Neuro-Fuzzy System which considers heterogeneous data and aids in the prediction of best suitable weather and soil status for improvising crop production. The section experiments details concerning the research as well as results discussions. The techniques of Co-EANFS and Map reduce focuses on accurately identifying weather prediction and effectively examining weather and soil for achieving crop production. Huge datasets are being employed in the work and various features provide enhanced and comparative outcome compared to the rest.

Table 1, presents throughout performance of Hadoop file system, comparing the output with rest of the available techniques like Hash algorithm, Collaborative filtering using Map reduce technique HFDS. It's revealed that the proposed technique of Map reduces yields improvised output compared to the rest others.

Table 4 The recommendation system

Recommendation rate	Error
96.45	3.55

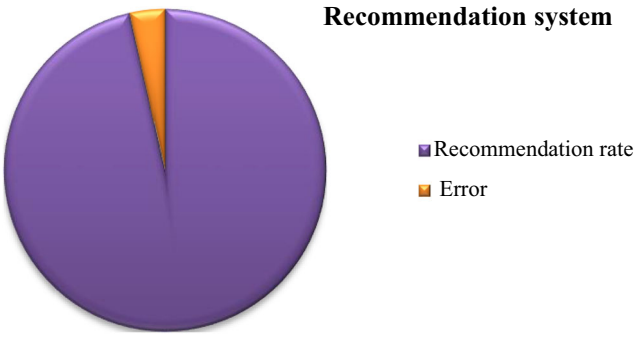


Fig. 7 Error performance

Table 2 presents comparison of various classification Techniques. Throughout performance and comparison results are being compared with various existing classification techniques like, RBFN, SVM, Co-EANFS and K-NN. It's revealed that the proposed Co-EANFS classification technique yield in improvised output than the existing classification system. Later, the study works upon detecting overall performance that includes efficiency, accuracy and processing time.

Figure 5, demonstrates the comparison involving various classification Techniques results. Classification technique's performance along with comparison output with prevailing techniques is being depicted in the graph. It's revealed that the proposed Co-EANFS classification technique yield in improvised output than the existing classification system.

Table 3, demonstrates the comparison with various available techniques such as Numerical Weather Prediction, C5, Deep learning Prediction and Association Rule Mining. The proposed technique of Association rule mining yields in high accuracy compared to rest others.

Figure 6 represents prediction techniques results concerning the weather prediction and demonstrating a comparison with various existing techniques like Deep Learning Prediction, C5.0, and Numerical Weather Prediction and Association rule mining. Here too it's revealed that the technique of Association Rule Mining generates effective output compared to rest of the Weather prediction approaches.

A complete Recommendation Rate and Error of the Recommendation System is depicted in Table 4.

Figure 7, presents the Error Performance related to recommendation system output concerning the best suitable crop for future recommendation to users. Here accuracy of 96.45% is achieved with error rate being 3.55%.

Conclusion

Weather Forecasting tends to be a technical and scientific concern over the years in the realm of climate prediction and dynamics theory all over. The research work is imbibed in big data scenario by enforcing Hadoop DFS and Co-EANFS for the purpose of classification and association rule mining for the purpose of Weather prediction. Also, methods namely, HDFS, classification and Prediction are being incorporated in the paper work. The Co-EANFS regularization is adopted for the purpose of implementation and for enhancing the recommender system's success ratio, the association rule mining using prediction is applied. Further in order to achieve high prediction accuracy, various regression model can be adopted. A comparative study amidst meteorological parameters pattern recognition and the smaller scale geographical region reveals better performance output and noteworthy accuracy in prediction.

Compliance with ethical standards

Conflict of interest This paper has not communicated anywhere till this moment, now only it is communicated to your esteemed journal for the publication with the knowledge of all co-authors.

Ethical approval This article does not contain any studies with human participants or animals performed by any of the authors.

References

1. Saraladevi, B., Pazhaniraja, N., Victer Paul, P., Saleem Basha, M. S., and Dhavachelvanc, P., Big data and Hadoop-a study in security perspective. In: *2nd International Symposium on Big Data and Cloud Computing*, 2015.
2. Alam, A., and Ahmed, J., Hadoop architecture and its issues. In: *International Conference on Computational Science and Computational Intelligence*, 2014.
3. Islam, N. S., Wasi-ur-Rahman, M., Xiaoyi, L., and Panda, D. K., Efficient data access strategies for Hadoop and spark on HPC cluster with heterogeneous storage. In: *IEEE International Conference On Big Data*.
4. Ogawa, H., Nakada, H., Takano, R., and Kudoh, T., SSS: An implementation of key-value store based MapReduce framework. In: *2nd IEEE International Conference on Cloud Computing Technology and Science*, 2010.
5. Kc, K., and Anyanwu, K., Scheduling Hadoop jobs to meet deadlines. In: *2nd IEEE International Conference on Cloud Computing Technology and Science*, 2010.
6. Sutariya Kapil, B., and Sowmya Kamath, S., Resource aware scheduling in Hadoop for heterogeneous workloads based on load estimation. In: *ICCCNT*, 2013.
7. Ghosh, S., Biswas, D., Biswas, S., (Sarkar), D. C., and Sarkar, P. P., Soil classification from large imagery databases using a neuro-fuzzy classifier. *Can. J. Electr. Comput. Eng.* 39:333–343, 2016.
8. Kuehnea, G., Llewellyna, R., Pannellb, D. J., Wilkinsonc, R., Dollingd, P., Ouzmana, J., and Ewinge, M., Predicting farmer uptake of new agricultural practices: A tool for research extension and policy, 115–125. © 2017 Published by Elsevier Ltd.
9. Wiston, M., and Mphale, K. M., Weather forecasting: From the early weather wizards to modern-day weather predictions. *Journal of Climatology & Weather Forecasting* 6(2):1–9, 2018.
10. Vipulkumar, S. D., and Somani, H., A survey on data placement in heterogeneous cloud environment for big data. 4(4):583–587. © 2016 IJEDR.
11. Abhishek, K., Kumar, A., Ranjan, R., and Kumar, S., A rainfall prediction model using artificial neural network. In: *2012 IEEE*

Control and System Graduate Research Colloquium (ICSGRC), 2012, 82–87.

12. Samya, R., and Rathipriya, R., Predictive analysis for weather prediction using data mining with ANN: A study. *International Journal of Computational Intelligence and Informatics* 6:150–154, 2016.
13. Kushwaha, A. K., and Bhattachrya, S., Crop yield prediction using agro algorithm in Hadoop. *IRACST - International Journal of Computer Science and Information Technology & Security (IJCSITS)* 5(2):271–274, 2015.
14. Ankalaki, S., Chandra, N., and Majumdar, J., Applying data mining approach and regression model to forecast annual yield of major crops in Different District of Karnataka. *International Journal of Advanced Research in Computer and Communication Engineering* 5(2):25–29, 2016.
15. Poongodi, S., and Babu, M. R., Prediction of crop production using improved C4.5 with ANFIS classifier. *International Journal of Control Theory and Applications* 10(21):121–132, 2017.
16. Ramesh, D., and Vardhan, B. V., Data mining techniques and applications to agricultural yield data. *International Journal of Advanced Research in Computer and Communication Engineering* 2(9):3477–3480, 2013.
17. Bhardwaj, P., and Singh, M., A review of different techniques utilized for-casting crop yield. *International Journal of Engineering & Technology* 7(2,8):268–270, 2018.
18. Dahikar, S., and Rode, S., Agricultural crop yield prediction using artificial neural network approach. *International Journal of Innovative Research in Electrical, Electronics, Instrumentation, and Control Engineering* 2(1):683–686, 2014.