International Journal of Information and Decision Sciences

ISSN online: 1756-7025 - ISSN print: 1756-7017
https://www.inderscience.com/ijids

Modelling big data analysis approach with multi-agent system for crop-yield prediction

Jaya Sinha, Shri Kant, Megha Saini

DOI: 10.1504/IJIDS.2023.10054771

Article History:

Received:	01 October 2020
Last revised:	04 January 2021
Accepted:	23 January 2021
Published online:	20 March 2023

Modelling big data analysis approach with multi-agent system for crop-yield prediction

Jaya Sinha*

Department of Computer Science and Engineering,
Galgotias College of Engineering and Technology,
Greater Noida, India
Email: jsjoysinha@gmail.com
*Corresponding author

Shri Kant

Research and Technology Development Centre,
Sharda University,
Greater Noida, India
Email: shri.kant@sharda.ac.in

Megha Saini

Department of Information Technology,
Galgotias College of Engineering and Technology,
Greater Noida, India
Email: meghasaini198@gmail.com

Abstract: Big data environment in current scenario is dealing with challenges in handling inherent complexity residing in the massive heterogeneous, multivariate and continuously evolving real-time data along with offline statistics. The role of big data analytics to analyse such a highly diverse data also plays a significant role in estimating predictive performance of a system. This paper thus aims at proposing an intelligent agent-based architecture that coordinates with big data analytics framework to model a system with an objective to improve the predictive performance of system by handling such diverse data. The paper also includes implementing predictive algorithm to predict crop yield in the agricultural domain. Various machine learning analytical tools have been used for data analysis to produce comprehensive and more accurate prediction using the proposed architecture.

Keywords: multi-agent system; MAS; big data; data acquisition; data analysis; data storage; machine learning; intelligent agents.

Reference to this paper should be made as follows: Sinha, J., Kant, S. and Saini, M. (2023) 'Modelling big data analysis approach with multi-agent system for crop-yield prediction', *Int. J. Information and Decision Sciences*, Vol. 15, No. 1, pp.27–45.

Biographical notes: Jaya Sinha is an Assistant Professor at the Galgotias College of Engineering and Technology, Greater Noida, India. She has done her MTech (Information Technology) from the GGSIP University, Delhi and received her PhD (Computer Science) from the Sharda University, Greater

Noida, India. She has an experience of more than 15 years in the field of teaching. She has published five research papers in peer reviewed international journals and conferences. She has research interest in the area of intelligent agents, agent-based software engineering, data mining and machine learning. She has a lifetime membership in the Indian Society for Technical Education (ISTE).

Shri Kant is a Professor at the Research and Technology Development Centre (RTDC) of Sharda University, India. He has more than 35 years experience in Defence Research and Development Organisation (DRDO), India as a Scientist, coordinator and the Director of a DRDO lab. During this period, he has guided a team of scientists working on pattern recognition and soft computing application to be used for cryptanalysis. He has received his PhD (Mathematics) from the Institute of Technology, Banaras Hindu University, India. His areas of interest are special functions, cryptology, pattern recognition, cluster analysis and data mining. He has published more than 70 research papers in international, national journals and conferences. He has received commendation certificates and scientist of the lab award for exhibiting the excellence in pattern recognition application to cryptology.

Megha Saini is a student of BTech with specialisation in Information Technology at the Galgotias College of Engineering and Technology. She has a keen interest in research and likes to participate in hackathons. She has worked on a project e-kaksha, which was selected for AICTE Chhatra Vishwakarma Award, 2018, India.

1 Introduction

The world has entered in an era of big data where big data does not only mean massive or huge amount of data but it also deals with other characteristics of data as continuous generation of high pace data, data having diverse forms and sources among others. As per The National Institute of Standards and Technology, the big data is defined to be characterised by data attributes volume, velocity, variety and variability, the V's of big data (NIST, 2015). These metrics are utilised to classify big data sets. Along with these metrics, the big data architecture should also be scalable enough to ensure efficiency in storage, manipulation and analysis of big data (NIST, 2015; Hu et al., 2014). Many other metrics has also defined for big data; one such has added value to volume, velocity, variety and variability rearing it as 5V's of big data (Shashaj et al., 2019).

The basis of big data analytics is the procedure retrieving large amount of data from various source and then converting it into knowledge useful in many domains (Hu et al., 2014). Big data processing is specified by processes that identify interesting hidden patterns, stores massive and heterogeneous data with an efficient storage infrastructure in addition to data analysis (NIST, 2015). Such environment poses challenge associated with big data analytics focusing on data interpretation, prediction, system modelling and simulation among others to be met by research communities. Also, the current trend has shown that not only offline but real-time online data also plays a major role in predictive analytics (Hu et al., 2014). Thereby, recent development in big data analytics demands new paradigms to be developed to mine and analyse massive data being both offline and real-time so as to substantially improve the decision making process. In an effort to

achieve such an objective, we have considered multi-agent paradigm with intelligent agents (IAs) to model the system architecture. IA is an entity that is reactive, proactive, autonomous, adaptable, intelligent, collaborative and knowledgeable in behaviour (Jennings and Wooldridge, 2000). As big data analytics has an advantage of discovering and analysing hidden patterns underlying massive data and IAs being autonomous and adaptive in nature becomes the most feasible choice for modelling big data analytics framework to enhance predictive power of the system. Multi-agent system (MAS) composed of IAs is a suitable choice as it offers an adaptive environment where dynamic modification may improve the decision making capability of the system (Sinha et al., 2018).

This work has been motivated by orthogonal features of IAs that have the capability to handle both offline and real-time dynamic data. This dynamic capability of IAs when combined with big data analysis approach presents a paradigm for handling real-time big data.

In this paper, we have attempted to combine this capability of big data analysis with IAs to model an adaptive predictive system for predicting crop yield in agriculture sector. Machine learning (ML) algorithms were used for experimentation, to evaluate the predictive power and performance of the proposed agent-based architecture (Tsai et al., 2015). The main contribution of the experimental work focuses on predicting crop yield specifically for Indian horticulture crops using multi-agent and big data analysis paradigm with ML tools, for which no significant work has been done till yet. In this study we are exploring two possibilities:

1 Potential of MAS in big data analysis environment.

2 Effectiveness of agent-based system in agriculture for crop-yield estimation.

Novelty of the work lies in presenting and simulating multi-agent architecture having big data analysis capability which has been successfully applied in the domain of agriculture for crop-yield prediction.

Section 2 of the paper presents recent literature review in the area of MAS for big data analytics and related recent work in agriculture domain. Section 3 proposes multi-agent big data analysis architecture and highlights the functions of each of its components. Section 4 demonstrates system modelling of exemplar case study on crop yield prediction and Section 5 presents experiment and result of simulation. Section 6 presents the conclusion with future pipeline work.

2 Related research

Multi-agent paradigm is gaining a lot of popularity in big data analytics due to the requirement of identifying new paradigms suitable for analysing real-time data that varies with time in a dynamic environment. In this section, we have discusses some related work where authors have proposed multi agent architecture for supporting dynamic big data analytics.

Twardowski and Ryzko (2014), have proposed architecture for big data processing in real-time using multi-agent approach based on Lambda architecture. Authors have applied the proposed architecture to implement recommendation system and concluded that it is a suitable approach to be applied in such applications which process both online

and offline datasets. In 2016, Belghache et al. have also explored the possibility of applying MAS to big data analytics. They proposed adaptive multi-agent system (AMAS) architecture to detect real-time data correlation continuously in a dynamic environment with the help of agents. The proposed architecture promises to achieve fast detection of real-time correlation of features, better context-learning and scaling up non-exponential system with efficient memory requirement in a complex and distributed environment. Though architecture proposes promising outcomes but performance evaluation and validation of proposed system are yet to be done tasks (Belghache et al., 2016). In the same year, Baert et al. (2016) have proposed MAS to optimise task reallocation for processing large datasets using negotiation. The proposed MAS help in analysing large dataset using MapReduce framework. The system is mainly characterised by Mapper and Reducer agent and the works focuses on implementing MapReduce framework without pre-processing data concluding that MAS is well suited to dynamic reallocation tasks.

Elaggoune et al. (2018) in their work has presented an IA-based MapReduce framework for smart data extraction in the field of smart healthcare. The proposed framework promises advantage of being low cost by avoiding storage space wastage and a better decision making process using agents. Eliminating data redundancy in a distributed environment, tackling noise filtration without data loss and implementation of the proposed framework in smart healthcare and others are some of the constraints of current work to be handled in future. Shashaj et al. (2019) have designed a multi-agent specialisation system which represents a distributive MAS environment to initiate development of ambient intelligence (AmI) applications. Authors have also discussed suitability of the proposed agent environment under big data processing scenario. The future work intends to apply the proposed environment on an application to optimise customer/citizen operations. Also, Elaggoune et al. (2020) have proposed a multi-fuzzy agent system for smart data extraction in big data environment. In order to show the effectiveness of the proposed approach smart wireless sensor network was simulated using fuzzy agents with an objective to filter only relevant data from available input data in big data environment. They study showed that the proposed system improved the network efficiency in terms of power consumption and lifetime.

Some recent work related to agriculture domain has also been studied to understand various phenomenon used. AlShahrani et al. (2017) have proposed a recognition system to classify healthy and rotten crops based on common attributes. Image processing and bag of features (BoF) methodologies were used for classification and statistical measurements. Least mean square error (LMSE) has been used to evaluation performance of image processing technique. Authors have also compared the result of BoF classification with results using measure LMSE for classifying crops as healthy and rotten. Rajinikanth et al. (2020) in their work have focused on recognising benchmark crop-weed (BCW) images. Main aim of authors was to identify the most suitable technique to recognise BCW images with best accuracy and precision. The work combines the spider monkey optimisation (SMO), Kapur's multi-thresholding along with watershed segmentation methods for extracting crop-weed regions in BCW images.

During literature survey, mainly two research gaps were identified and have been aimed at in our work. Although, many research work focus on advantages of MAS and big data analytics but there is a lack of applicability of such architecture in the field of agriculture. Also, survey work related to phenomenon used within agriculture domain shows that no significant work has been done to capture real-time agricultural data using IAs with big data analysis.

Thereby, in a quest to identify new approaches to model big data processing tasks, these research work supports multi-agent paradigm to be suitable one. We have thereby found an opportunity to propose and simulate a multi-agent based architecture for big data analysis in agriculture domain. In addition to this, the possibility of applying ML techniques in our work has also been identified to further improve the system's predictive performance during analysis.

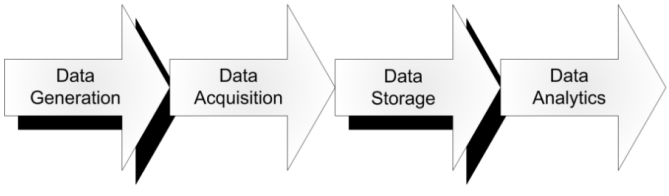
3 Multi-agent big data analysis architecture

Motivated by the successful projection of multi-agent paradigm at different stages of big data processing and its applicability in diverse domain, we have proposed an IA-based architecture to model an adaptive predictive system to assist data analysis process using ML tools. The paper also presents simulation of the proposed architecture for predicting crop yield in agriculture sector thereby contributing in precision agriculture.

3.1 Function and behaviour of IAs

In system engineering perspective, big data system has been presented as a value-chain consisting of four distinct consecutive phases shown in Figure 1. Data generation, data acquisition, data storage and data analytics are the four major phases of big data life cycle (Hu et al., 2014).

Figure 1 Big data value chain



Source: Hu et al. (2014)

Inspired by this visualisation of whole big data system as a value-chain, the two phases data acquisition and data analytics has been modelled and simulated using IAs in the proposed MAS. Thereby, the proposed architecture is composed of two broad categories of adaptive IAs data acquisition agent (DAAgent) and data processing agent (DPAgent). The MAS architecture for big data analysis has been presented in Figure 2.

3.1.1 Data acquisition agent

As DAAgent simulates the functionalities of data acquisition process of big data value-chain, there are three major functions that are performed by DAAgent are (Hu et al., 2014; Di Martino et al., 2014):

- Data collection that includes gathering structured and unstructured data.
- Data pre-processing for cleaning and filtering of gathered data.
- Data transmission.

Figure 2 Proposed big data – agent architecture

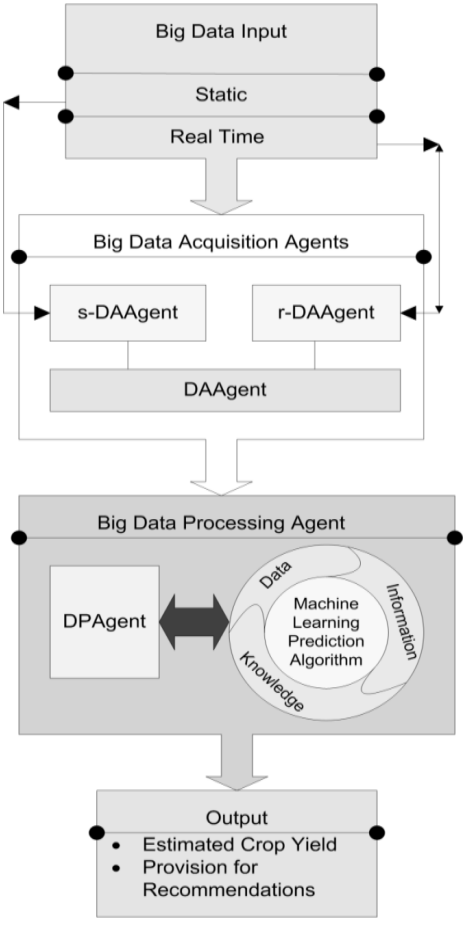
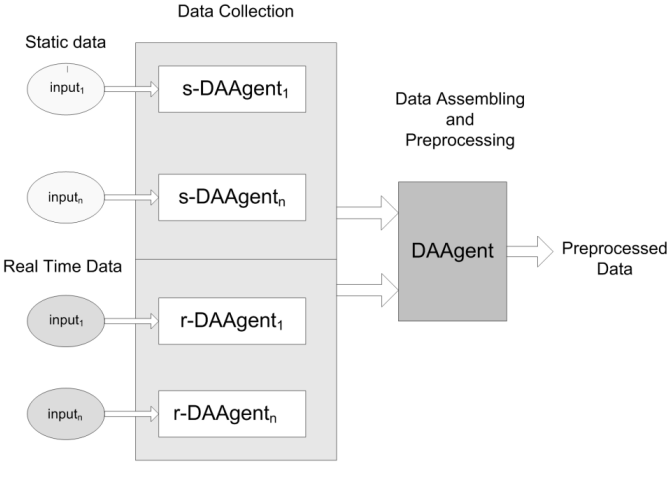


Figure 3 Big data acquisition using IAs



DAAgent is an IA that captures both form of data being either static or dynamic in nature with its two variants s-DAAgent (static) and r-DAAgent (real-time) presented in Figure 3.

s-DAAgent gathers those characteristic input feature set that remains static. Static data is either statistically computed or historic whose value does not change. Also, with an aim to capture the dynamic data which is more real-time in nature that keeps on changing frequently, the architecture also has an IA namely r-DAAgent. r-DAAgent has the responsibility of capturing data that keeps on changing at frequent intervals and interfaces with sensors in a dynamic environment. The time interval depends on the type of data collected for specific application to be implemented where small intervals may be in some hours, days or weeks.

DAAgent also act as a coordinator agent to manage data collection. After the data is gathered, DAAgent pre-processes and extracts the relevant data attributes for experimental dataset. Pre-processing includes removing missing values, outlier detection, its removal and partitioning the data into training and test sets. After normalising the data sets, pre-processed data is transmitted to be stored either in big data storage using Hadoop or cloud-based virtual storage in cloud (Hu et al., 2014; Elaggoune et al., 2018; Geng et al., 2019).

3.1.2 Big data storage

The next phase after data acquisition phase in big data value chain is data storage. In this phase, the massive data collected is stored in a defined format using data management framework. The data is managed using file systems such as Hadoop distributed file storage system (HDFS), Hive and others, database technologies such as NoSQL along with some frameworks as MapReduce, Spark and others. Organised data allows for easy analysis and extraction of hidden knowledge in the form of correlation among collected datasets (Hu et al., 2014; Elaggoune et al., 2018).

For simplicity, the structure and function of the storage model has not been discussed and its elaboration is beyond the scope of this study and gives a future work direction. This study keenly focuses on big data acquisition and analysis phase along with recommending the proposed architecture suitable for agricultural sector.

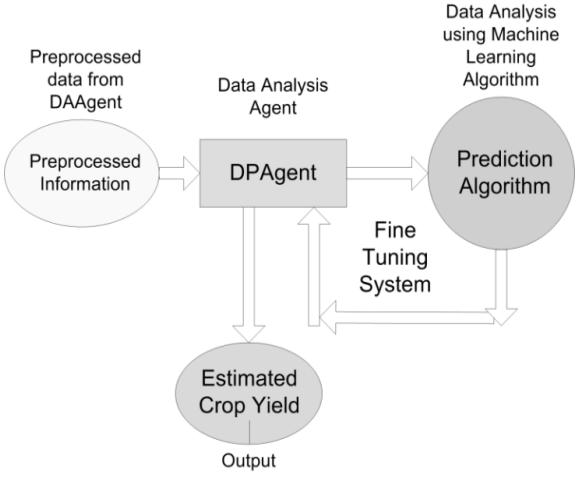
3.1.3 Data processing agent

The responsibility of DPAgent is to perform predictive data analysis and interpretation which are the main features of data analysis in big data value chain (Hu et al., 2014). Although, there are various tools and techniques that can be applied on big data for its analysis, ML has been one such technique that gives an opportunity to develop more robust system for comprehensive prediction and perception of complex data in diverse areas (Kersting and Meyer, 2018; Liakos, 2018). The predictive analytics by DPAgent utilises ML predictive algorithms to identify the characteristic patterns or correlation among underlying the massive input datasets during analysis.

The DPAgent thus analyses and interprets the stored filtered data by deploying different ML algorithms during data analysis. The knowledge extracted or insights thereby observed by the system recommend set of actions that helps in optimising system's predictive performance. As shown in Figure 4, DPAgent receives input from DAAgent for analysis. Since, IAs have the capability to react in a dynamic environment,

the DPAgent will act autonomously as and when there is a significant change in the real-time environment variables or characteristic feature of input dataset supplied by DAAgent. This implies that the ML algorithm will be executing at frequent intervals by incorporating changed values provided by DAAgent. This continuous computation by processing both static and real-time dynamic data makes DPAgent adaptive in nature and contributes towards optimising the overall prediction process, making the whole system adaptive. The data interpretations thereby result in identifying correlation among underlying data segments facilitating decision making.

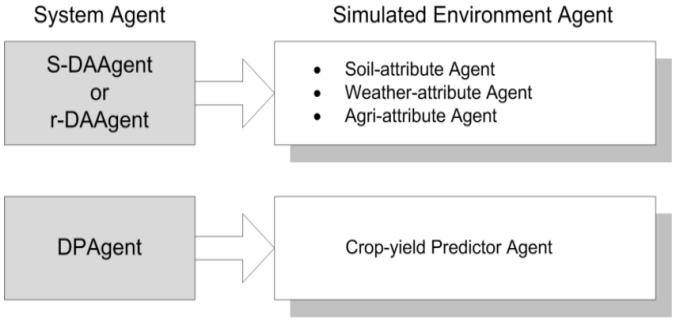
Figure 4 Big data processing and analysis using IAs



4 Exemplar case study: crop yield prediction

Agriculture in India is one of the largest employment sectors, contributing major percentage in Indian economy. Also, India being a land of diverse weather and soil, predicting crop yield is one of the major challenges in agriculture farming sector and despite development in technology still many of the farmers are following traditional techniques in agriculture farming (Shastry and Sanjay, 2020). This quest for equipping traditional practices with modern technology opens up many avenues of research in this domain. This paper emphasises on utilising MAS and big data analysis with ML tools to model and automate crop yield prediction process (Liakos, 2018). The productivity of crop is dependent on different attributes which can be classified as soil, weather and other agricultural attributes. Some key soil attributes include soil moisture content, surface temperature, soil organic matter, type, pH value and depth. Key weather attributes includes rainfall, temperature, humidity measure and solar radiation. Uncertainty in weather conditions makes most of the weather attributes to be dynamic in nature. Farming area, amount of fertiliser application, crop species, date of sowing, seasonal duration cultivar (SDR), irrigation water depth are some of the important other agricultural attributes for estimating crop yield (Gonzalez-Sanchez et al., 2014; Majumdar et al., 2017).

Figure 5 Simulated agents for exemplar environment



In order to handle such multivariate, heterogeneous and continuously evolving real-time data that affects crop productivity MAS has been modelled for agriculture big data analysis. Figure 5 demonstrates that in the proposed architecture for the exemplar problem domain soil-attribute agent, weather-attribute agent and agri-attribute agent simulating functionality of DAAgents whereas crop-yield predictor agent functions as DPAgent.

4.1 DAAgent (soil-attribute agent, weather-attribute agent and agri-attribute agent)

DAAgent simulates soil-attribute agent, weather-attribute agent and agri-attribute agent for the exemplar case study. r-variant of each DAAgents captures corresponding real-time data if any whereas s-variant gathers static or fixed value data. Its main functionality includes data collection, pre-processing and transmitting dataset to be processed by DPAgent. All the DAAgents collects the respective input data as sensor data for soil, satellite data for weather, online published statistics among others from various sources (Kamilaris et al., 2017). All the collected observations are further compiled as dataset.

4.2 DPAgent (crop-yield predictor agent)

Crop-yield predictor agent being application specific agent simulates the functionality of DPAgent. After analysing stored data using different data analytic tools and techniques as per the proposed algorithm in Figure 6, crop yield estimation is generated as output. This information gives an estimation insight to farmers much before the crop harvesting time thereby enables farmers to apply various preventive measures in the mean time to increase the crop yield, if estimated crop yield is low.

5 Experiment and result analysis

5.1 Description of input feature data set

For predicting crop yield, the dataset for experimentation consists of six features pertaining to weather, soil and other agricultural factors presented in Table 1.

The experimental data was collected for 27 states and four union territories (UT) of India to estimate productivity of horticulture crops for the year 2018, from different sources mainly IMD and Government of India for weather data as mean rainfall (mm) for the sates and UT, Department of Agriculture, Cooperation and Farmers Welfare, Government of India for crop related data along with others (Bhattacharyya and Pal, 2015) for soil data as soil type and depth range (cm) for each soil type. Horticulture crop statistics has been collected from website of Department of Agriculture, Cooperation and Farmers Welfare, Government of India and seven different crop types have been chosen to estimate crop yield. This includes fruits, vegetables, plantation, spices, aromatic, flower_loose and flower_cut and collected data as area (`000 Ha) and past crop productivity (`000 MT) from 2011 to 2018 and all the collected data information maintained in the dataset has been shown in Table 1. There are in all 1,305 observations after data pre-processing in the crop dataset for estimation.

Table 1 Input and output determinants for the test case simulation

	Selected features for experimental dataset
1	Mean rainfall (mm)
2	Area (`000 Ha)
3	Past crop productivity (`000 MT) from 2011 to 2018
4	Soil type (mountain, laterite, grey and brown, red, black, desert, alluvial)
5	Depth range (cm) for each soil type (SDP1:below 25, SDP2: 25, SDP3: 25–50, SDP4: 100–300, SDP5: below 300, SDP6: above 300)
6	pH range for each soil types (5.0–6.5, below 5.5, 7.6–8.9, 5.2–7.5, 6.5–8.4, 7.6–8.4)

5.2 Application of ML predictive algorithms

ML algorithms have striking capability to learn patterns inside data automatically, so in this paper, to evaluate performance of system.

5.2.1 Linear regression

Linear regression is statistical modelling technique which identifies and establishes a linear relationship between a set of independent or explanatory variables A, where $A \equiv \{A_1, A_2, \ldots A_n\}$ and a dependent variable B by fitting a linear equation of the form given in equation (1) during training.

$$B_i = \beta_0 + \beta_1 A_{i1} + \beta_2 A_{i2} + \ldots + \beta_n A_i \tag{1}$$

While testing it predicts the value of B for a new input set as given in equation (2).

$$B' = \beta A + \epsilon \tag{2}$$

where ϵ is the residual error component during linear fit (Gonzalez-Sanchez et al., 2014).

We have applied multiple linear regression (MLR) to build a linear model by establishing linear relationship between various factors responsible for affecting crop production to predict future crop yield.

5.2.2 Instance-based k

In Waikato Environment for Knowledge Analysis (WEKA), instance-based k (IBk) is a form of supervised k-nearest neighbours technique and the method predicts class of a new target sample by identifying *k* closest or most similar samples from the training. To make predictions for regression problems, weighted mean of the *k* closest similar samples are taken. The *k* value which controls the neighbourhood is challenging to estimate (Mucherino et al., 2009). The value of *K* for experimentation was chosen to be 5 and 7.

5.2.3 Decision tree regression using REPTree

Decision trees are the simplest ML technique with a tree like structure. It identifies and chooses the best input feature as root node of the tree and the tree grows incrementally by partitioning the dataset on different input features to reach target node or leaf node. In this experimentation using WEKA, REPTree method has been used which generates multiple trees during different iterations using the decision tree logic and then from all the trees generated, an optimal one is selected (Gonzalez-Sanchez et al., 2014).

5.2.4 Support vector regression

Support vector regression (SVR) uses the principle of support vector machine (SVM) learning which fits a linear boundary to clearly separate samples of two classes. The boundary is established by identifying the largest distance between near samples of the classes to be classified. SVR uses kernel functions for linear regression to map sample points in low dimension into high dimensional space and fits hyperplane to separate samples of two classes (Awad and Khanna, 2015). The experimentation utilises sequential minimal optimisation learning algorithm (SMOReg) in WEKA.

5.2.5 Regression with multi-layer perceptron

Multi-layer perceptron (MLP) is a feed forward artificial neural network (ANN) that can be suitably applied to regression problems in real world order to predict a real valued target. MLP has three layers input, hidden and output with activation functions and uses backpropagation algorithm for training the network which mainly involves adjusting the weight vector corresponding to given output. Each neuron in the input layer corresponds to one input feature and neuron in the output layer gives predicted output, in this work it is crop yield estimation. ANN is an efficient technique to process uncertain, complex and nonlinear data (Gonzalez-Sanchez et al., 2014; Liakos, 2018).

After testing of the system is done, estimation accuracy of each algorithm has been analysed on computed estimated crop yield for the year 2018 using various statistical error metrics. After performance efficiency of each ML algorithm is compared subsequently suitable method can be recommended.

5.3 Predictive algorithm

Algorithm for prediction has been described in Figure 6. The proposed algorithm was implemented using following ML regression techniques with help of WEKA tool (Frank

et al., 2016). WEKA has been applied for simulating the exemplar case because of its popularity in providing various ML algorithms with user friendly visualisation tools.

Figure 6 Proposed algorithm

Predict_Crop_Yield: Computes estimated crop yield B_e for each crop type
Begin
Input: A= {{P}, {W}, {S}}
P= {P_a, C_a, Y_r, S_n}
W= {R_m}
S= {S_t, S_d, S_{ph}}
B= {C_a}
Output: B_e= {C_e}
where,
P_a- Planting area, C_a- Past actual crop yield, Y_r- Year of crop cultivation
S_n,- Indian State/Union Territory, R_m,- Mean rainfall, S_t,- Soil type,
S_d,- Range of soil depth, S_{ph},- Soil ph range, C_e- Estimated crop yield
Data collection involves gathering both static and dynamic data
Data_Collection
{
s-Weather-attribute Agent.Action()
{W= get_ Weather_Attribute( )}
s-Soil-attribute Agent.Action()
{S= get_ Soil_Attribute( )}
s-Agri-attribute Agent.Action()
{P= get_ Agri_Attribute( )}
}
Real_Time_Data_Collection
{ while (true)
{r-Weather-attribute Agent.Action()
{ W_1= get_ Weather_Attribute_Sensor( )}
r-Soil-attribute Agent.Action()
{ S_1= get_ Soil_Attribute_Sensor( )}
}
Set Timer at frequent time interval
}
Data set creation and preprocessing of collected data.
Subsequent dataset modification based on real time data captured at frequent intervals.
Create_Data_Set
{
DAAgent .Action()
{ Set A1= {{P}, {W}, {S}, {W_1}, {S_1}}
Identify driver attributes and modify dataset
Set A= {P_a, C_a, R_m, S_t, S_d, S_{ph}}
}
Data transformation and analysis using Regression as ML technique that involves training and testing the modeled system.
Data_TA
{ DPAgent .Action() {
ML_RegressionAlgo (I)
{ Train system to identify mapping f(A)=B
Iterate for n times where n=no. of years/epochs to fine tune syster
Compute Error, ϵ = B-B_e
return (B_e=f(A))
}
}
Repeat for N techniques until system outputs reasonable result
}
End

ML algorithms are used with an objective to identify a function f for mapping a given input with corresponding output as:

$$f(A) = B.$$

This algorithm is expected to train the system repeatedly for n observations as training dataset until error minimises. The training dataset contains 70%, validation dataset contains 15% and test dataset contains 15% of the total collected observations.

The proposed algorithm in Figure 6 demonstrates the procedure of implementation and working of IAs. Features mentioned in Table 1 act as input data, and the algorithm will return estimated crop yield (Be). 'Data_Collection' and 'Real_Time_Data_Collection' functions automates the data gathering process and captures both real-time and offline input data either through various sensors or from data stores using IAs. s-Soil-attribute Agent, s-Weather-attribute Agent and s-Agri-attribute Agent captures offline data respectively for soil (soil type, soil depth, soil ph), weather (mean rainfall) and other agricultural attributes (planting area, past actual crop yield, Indian states or UT, year of cultivation). r-Soil-attribute Agent, r-Weather-attribute Agent captures real-time evolving data for weather as changes in mean rainfall and soil sensor as changes in soil ph.

DAAgent there continuously interact with the system environment by identifying the type of input and accordingly modifies the dataset as soon as significant changes are observed in real-time data described in function 'Create_Data_Set'.

5.4 Qualitative assessment

In terms of analysing qualitative performance of the model's behaviour and design aspects, the proposed MAS design architecture and its simulation shows that the proposed system promises to be highly autonomous and adaptive in nature.

5.4.1 Autonomy

The functions 'Data_Collection' and 'Real_Time_Data_Collection' of the proposed algorithm in Figure 6 shows that, IAs being autonomous can automate the data gathering process for both real-time and offline data either through various sensors or from data stores. Also, to process offline data, IAs play a significant role in capturing a variety of real-time evolving data from different sources such as weather and soil sensors and keep on updating the data store as demonstrated in function 'Create_Data_Set' of proposed algorithm.

5.4.2 Adaptive

Along with being autonomous, adaptive, cooperative and collaborative nature of IAs allows them to interact and work with each other by capturing and incorporating the changes in the model, thereby making a dynamic, adaptable and robust system for big data analysis environment. Also, as described in function 'Data_TA' of proposed algorithm in Figure 6 since DPAgent processes both static and real-time data received by DAAgent, thereby termed as adaptive in nature. This adaptive nature of DAAgent and DPAgent contributes in designing adaptive and robust agent architecture as DAAgent continuously interact with the system environment by identifying the type of input and

accordingly DPAgent changes the processing functionality of system which has been diagrammatically shown in Figures 3 and 4 and the functionality of both agents has been described in sub-Sections 3.1.1 and 3.1.3.

5.5 Quantitative assessment

As root mean squared error (RMSE) and mean absolute error (MAE) are the most commonly used standard statistical error metric, our analysis majorly attributes model's performance on these two. Other measures are also taken into consideration as correlation coefficient (R), relative absolute error (RAE), root relative square error (RRSE) as each of these metrics help evaluate model's performance capturing different statistical aspects.

Tables 2 and 3 respectively stores the results obtained during training and testing for chosen ML algorithms and best computed values are marked as bold. As R signifies the strength of linear relationship of the model, the algorithm that gives the maximum value is best suitable one. Table 2 clearly shows that algorithm IBK has the maximum value of R (0.883) at $K = 5$ whereas during testing Table 3 demonstrates that the algorithm IBK has the highest value of R (0.961) at $K = 7$ and R (0.942) at $K = 5$ from rest of the algorithms. Thus, choosing R as performance metric, test data shows that IBK gives better correlation with a close competition by MLR and RepTree both having R value (0.892) followed by SMOReg with R value (0.873) and lastly MLP with R (0.589).

As RMSE gives the amount of spread of residual or prediction error and MAE gives an average of all absolute errors computed on all instances of the data sets, thus smaller the value of these metrics corresponds to better accuracy in prediction performance of the model. From Table 2, which stores RMSE and MAE values obtained during training, it is clear that the algorithm IBK at $K = 7$ has the smallest value of RMSE (1.518) and MAE has the smallest value for IBK at $k = 5$ with MAE (5.570. For the test set errors, Table 3 clearly demonstrates that the IBK has the smallest value of RMSE (1.584) and smallest value of MAE (5.854) at $K = 7$ followed by RMSE (1.72), MAE (6.318) at $K= 5$ from rest of the algorithms. Thereby, choosing RMSE and MAE as performance metric, IBK gives better predictive performance. For RMSE the next algorithm is RepTree (2.07), MLR (2.094), SMOReg (2.357), MLP (3.059) in sequence and for MAE the next algorithm is RepTree (6.97), SMOReg (10.895), MLR (13.768), MLP (15.516).

Table 2 Computed performance measures during training with ten cross-fold validation

Performance measure	*MLR*	*IBK*		*RepTree*	*SMOReg*	*MLP*
		K = 5	*K = 7*			
R	0.794	*0.883*	0.869	0.839	0.874	0.874
MAE	14.307	*5.570*	7.103	6.426	9.314	9.314
RMSE	2.389	1.84	*1.518*	2.139	2.091	2.091
RAE	59.498	*23.167*	29.54	26.72	38.756	38.756
RRSE	60.799	*46.816*	49.65	54.414	53.183	53.183

Figure 7 Computed performance values during testing for each metric R, MAE, RMSE, RAE and RRSE for the algorithms MLR, IBK (K = 5), IBK (K = 7), RepTree, SMOReg, MLP

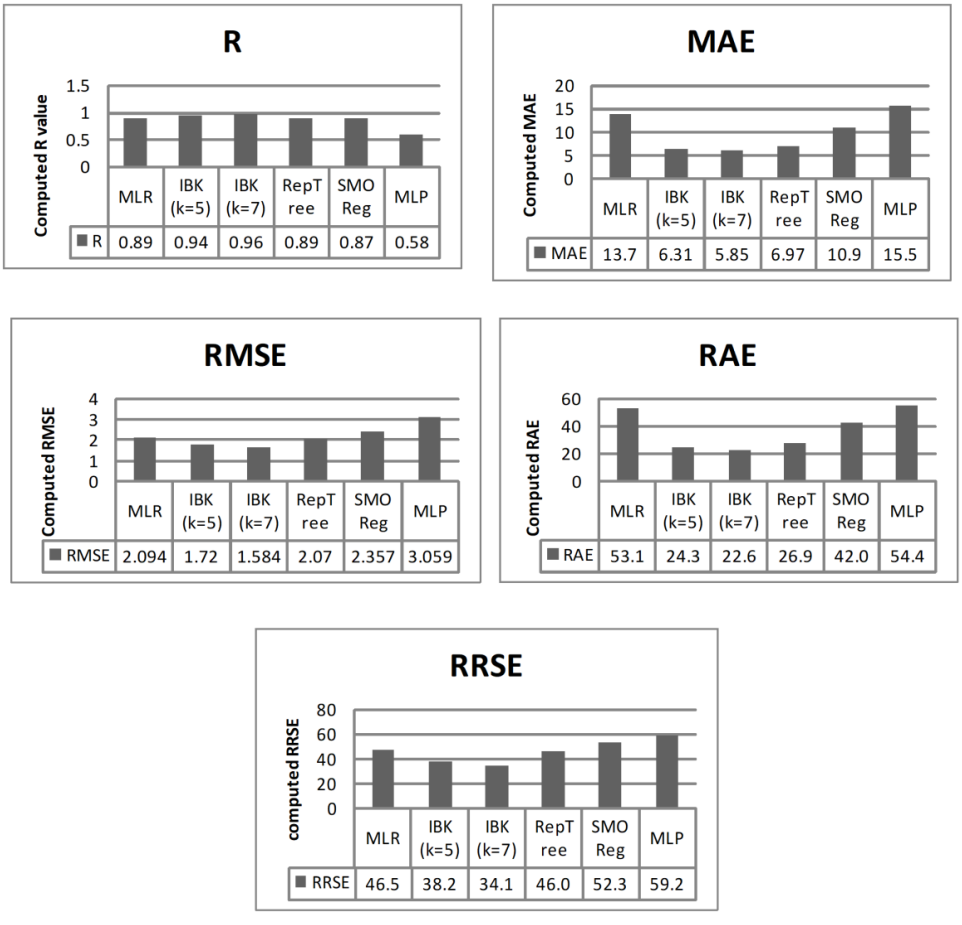
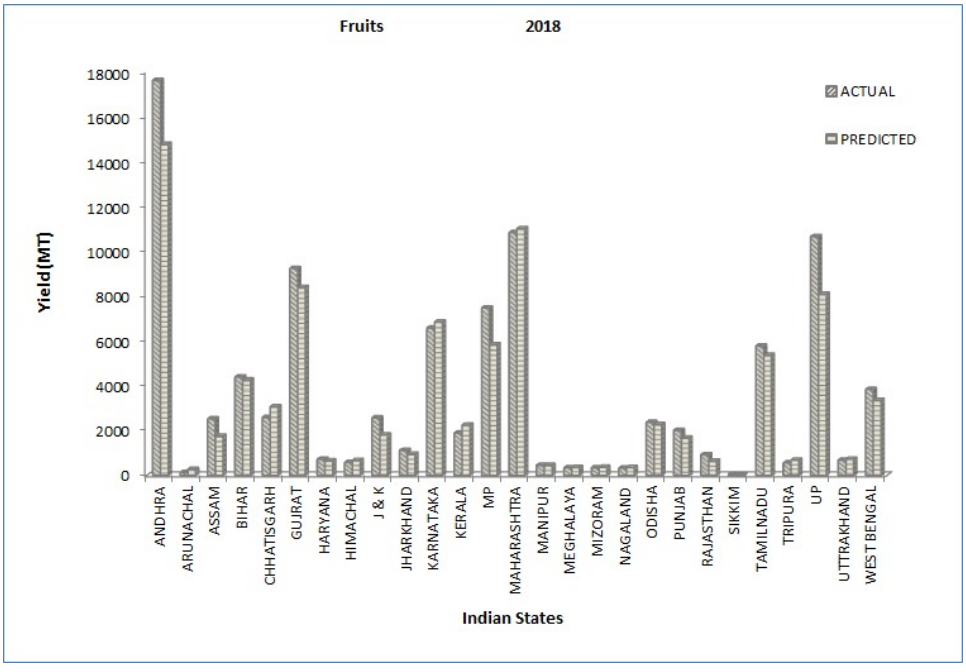


Table 3 Computed performance measures on test dataset

Performance measure	*MLR*	*IBK*		*RepTree*	*SMOReg*	*MLP*
		K = 5	*K = 7*			
R	0.892	0.942	*0.961*	0.892	0.873	0.589
MAE	13.768	6.318	*5.854*	6.97	10.895	15.516
RMSE	2.094	1.72	*1.584*	2.07	2.357	3.059
RAE	53.145	24.387	*22.596*	26.904	42.057	54.472
RRSE	46.532	38.23	*34.193*	46.019	52.388	59.21

Figure 8 Actual and predicted crop yield for the crop type 'fruits' using IBK (k = 7) for the year 2018



Using RAE and RRSE where value 0 signifies an ideal perfect fit for the model so best algorithm will be decided based on the smallest value of these metrics. Table 2 shows that the algorithm IBK at k = 5 has the smallest RAE (23.167) and RRSE (46.816) during training. For the test set, Table 3 clearly demonstrates that the IBK has the smallest value of RAE (22.596) and RRSE (34.193) at K = 7 followed by RAE (24.387), RRSE (38.23) at K = 5 as compared to rest of the algorithms. Thus, choosing RAE and RRSE as performance metric, IBK gives better performance closely followed by RepTree with RAE (26.904), RRSE (46.019). For RAE the next algorithm is SMOReg (42.057), MLR (53.145) and lastly MLP (54.472) whereas for RRSE the next algorithm is MLR (46.532), SMOReg (52.388) followed by MLP (59.21).

All the obtained results have been summarised and can be visually interpreted from Figure 7. Overall, on an average based on the results computed by experimentation, Figure 7 clearly shows that, the algorithm's performance can be ordered as IBK giving overall better predictions, i.e., highest R, smallest MAE, RMSE, RAE and RRSE when compared to rest of the others. In such ordered sequence IBK is being followed by RepTree, MLR, SMOReg and MLP.

Although, the experimental work uses seven horticulture crop types to estimate crop yield for the year 2018 mentioned in Section 5.1, Figure 8 presents the result obtained for one of the major horticulture crop type 'Fruits' only to present an insight of the experimental work done similarly for other crop types. Figure 8 demonstrates result of simulation obtained for 27 states as mentioned in Section 5.1 and shows the bar graph of the actual and estimated crop yield of each state for the year 2018 for 'fruits' using only IBK at K = 7. The statistical results presented in Table 3 which has been estimated for all

seven crop types during experimentation, shows that IBK algorithm has been best predictive algorithm as compared with others when applied on given input data'.

6 Conclusions

The objective of the proposed work was twofold. To demonstrate the applicability of MAS to model big data analysis architecture in order to process data that is dynamic in nature. Study shows that there exists a lot of scope to improve the performance of such an architecture using IAs, as their autonomous and adaptive nature helps to capture and process heterogeneous data that is both real-time continuously evolving stream and offline. Experimentation also shows that such an IA based data driven dynamic model can be successfully applied to predict crop yield in agriculture sector using ML algorithms and thereby is suitable for automated decision recommendations. Such predictions will be definitely helpful to farmers for planning and executing various preventive measures to improve the productivity of crop in case of low estimated crop yield than expected.

Future work will focus on further improving the performance of the model to predict with better accuracy by incorporating real-time variables due to changing soil and weather conditions which is also one of the challenges and the limitation of the current work. In addition to this, it is also challenging to estimate the performance of the model in situation of global pandemic as world is facing currently due to spread of COVID 19 viral infection where other socio-economic factors will have to be identified affecting the crop yield.

References

AlShahrani, A.M., Al-Abadi, M.A., Al-Malki, A.S., Ashour, A.S. and Dey, N. (2017) 'Automated system for crops recognition and classification', in Dey, N., Ashour, A. and Acharjee, S. (Eds.): *Applied Video Processing in Surveillance and Monitoring Systems*, pp.54–69, IGI Global, DOI: 10.4018/978-1-5225-1022-2.ch003.

Awad, M. and Khanna, R. (2015) 'Support vector regression', *Efficient Learning Machines*, pp.67–80, Apress, Berkeley, CA, DOI: https://doi.org/10.1007/978-1-4302-5990-9_4.

Baert, Q., Caron, A., Morge, M. and Routier, J-C. (2016) 'Fair multi-agent task allocation for large data sets analysis', in *Proceedings of PAAMS 2016 – 14th International Conference on Practical Applications of Agents and Multi-Agent Systems*, Sevilla, Spain, p.12.

Belghache, E., Georgé, J. and Gleizes, M. (2016) 'Towards an adaptive multi-agent system for dynamic big data analytics', in *Proceedings of International IEEE Conferences on Ubiquitous Intelligence and Computing, Advanced and Trusted Computing, Scalable Computing and Communications, Cloud and Big Data Computing, Internet of People, and Smart World Congress (UIC/ATC/ScalCom/CBDCom/IoP/SmartWorld)*, Toulouse, pp.753–758.

Bhattacharyya, T. and Pal, D. (2015) *State of Indian Soils'. State of Indian Agriculture – Soil*, 1st ed., pp.6–35, National Academy of Agricultural Sciences, New Delhi.

Department of Agriculture, Cooperation and Farmers Welfare, Government of India [online] https://agricoop.nic.in/ (accessed 2 September 2020).

Di Martino, B. et al. (2014) 'Big data (lost) in the cloud', *International Journal of Big Data Intelligence*, Vol. 1, Nos. 1–2, pp.3–17.

Elaggoune Z., Maamri R. and Boussebough I. (2018) 'A multi-agent framework for medical diagnosis driven smart data in a big data environment', in Auer, M. and Tsiatsos, T. (Eds.): *Interactive Mobile Communication Technologies and Learning, IMCL 2017, Advances in Intelligent Systems and Computing*, Vol. 725, Springer, Cham., https://doi.org/10.1007/978-3-319-75175-7_71.

Elaggoune, Z., Maamri, R. and Boussebough, I. (2020) 'A fuzzy agent approach for smart data extraction in big data environments', *Journal of King Saud University – Computer and Information Sciences*, Vol. 32, No. 4, pp.465–478.

Frank, E., Hall, M.A. and Witten, I.H. (2016) *The WEKA Workbench. Online Appendix for Data Mining: Practical Machine Learning Tools and Techniques*, 4th ed., Morgan Kaufmann [online] https://www.cs.waikato.ac.nz/ml/weka/Witten_et_al_2016_appendix.pdf (accessed 7 May 2020).

Geng, D. et al. (2019) 'Big data-based improved data acquisition and storage system for designing industrial data platform', *IEEE Access*, Vol. 7, pp. 44574–44582 [online] https://ieeexplore.ieee.org/document/8681030.

Gonzalez-Sanchez, A., Frausto-Solis, J., and Ojeda-Bustamante, W. (2014) 'Predictive ability of machine learning methods for massive crop yield prediction', *Spanish Journal of Agricultural Research*, Vol. 12, No. 2, p.313.

Hu, H., Wen, Y., Chua, T. and Li, X. (2014) 'Toward scalable systems for big data analytics: a technology tutorial', *IEEE Access*, Vol. 2, pp.652–687 [online] https://ieeexplore.ieee.org/document/6842585; doi: 10.1109/ACCESS.2014.2332453.

Indian Meterological Department (IMD) [online] https://mausam.imd.gov.in/ (accessed 5 May 2020).

Jennings, N.R. and Wooldridge, M. (2000) 'Agent-oriented software engineering', *Artificial Intelligence*, Vol. 117, No. 2, pp.277–296.

Kamilaris, A., Kartakoullis, A. and Prenafeta-Bold, F.X. (2017) 'A review on the practice of big data analysis in agriculture', *Computers and Electronics in Agriculture*, Vol. 143, No. C, pp.23–37.

Kersting, K. and Meyer, U. (2018) 'From big data to big artificial intelligence?', *Künstl Intell*, Vol. 32, pp.3–8, Springer [online] https://link.springer.com/article/10.1007%2Fs13218-017-0523-7#citeas; https://doi.org/10.1007/s13218-017-0523-7.

Liakos, K.G. (2018) 'Machine learning in agriculture: a review', *Sensors*, Vol. 18, No. 8, p.2674.

Majumdar, J., Naraseeyappa, S., and Ankalaki, S. (2017) 'Analysis of agriculture data using data mining techniques: application of big data', *Journal of Big Data*, Vol. 4, No. 1, pp.1–15.

Mucherino, A., Papajorgji, P.J. and Pardalos, P.M. (2009) 'k-nearest neighbor classification', *Data Mining in Agriculture. Springer Optimization and Its Applications*, Vol. 34, Springer, New York.

National Institute of Standards and Technology (NIST) (2015) *NIST Big Data Interoperability Framework: Volume 1, Definitions (Special Publication 1500-1)* [online] https://dx.doi.org/10.6028/NIST.SP.1500-1 (accessed 24 April 2020).

Rajinikanth, V., Dey, N., Satapathy, S.C. and Kamalanand, K. (2020) 'Inspection of crop-weed image database using Kapur's entropy and spider monkey optimization', in Das, K., Bansal, J., Deep, K., Nagar, A., Pathipooranam, P. and Naidu, R. (Eds.): *Soft Computing for Problem Solving. Advances in Intelligent Systems and Computing*, Vol. 1048, pp.405–414, Springer, Singapore.

Shashaj, A. et al. (2019) 'A distributed multi-agent system (MAS) application for continuous and integrated big data processing', in *Proceedings of 15th European Conference on Ambient Intelligence (AmI 2019)*, Springer, pp.350–356.

Shastry, A.K. and Sanjay H. (2020) 'Data analysis and prediction using big data analytics in agriculture', in Pattnaik, P. et al. (Eds.): *Internet of Things and Analytics for Agriculture*, Vol. 2, pp.201–224, Springer, Singapore.

Sinha, J., Ravulakollu, K. and Kant, S. (2018) 'Software development approaches significant for runtime software evolution: a review', in *Proceedings of IEEE International Conference on Advances in Computing, Communication Control and Networking (ICACCCN)*, pp.541–545.

Tsai, C.W. et al. (2015) 'Big data analytics: a survey', *Journal of Big Data*, Vol. 2, Article 21, pp.1–32, DOI 10.1186/s40537-015-0030-3.

Twardowski, B. and Ryzko, D. (2014) 'Multi-agent architecture for real-time big data processing', in *Proceedings of IEEE/WIC/ACM Int. Joint Conf. on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)*, Warsaw, pp.333–337.