Ann. Agric. Res. New Series Vol. 39 (2) : 210-215 (2018)

Role of big data in agriculture- A statistical prospective

Prakash Kumar[1], Anil Kumar[1], Sanjeev Panwar[2], Sukanta Dash[1], Kanchan Sinha , Vipin Kumar Chaudhary and Mrinmoy Ray[1]

ICAR-Indian Agricultural Statistics Research Institute
e-mail: prakash289111@gmail.com

Received : August 2017 ; Revised Accepted: January 2018

ABSTRACT

Data are playing an important role making good planning and policies for agricultural growth and development. Population growth and climate change are worldwide trends that are increasing the importance of using big data science to improve agriculture. Add to that land degradation increasing marginal land and loss of biodiversity are better deals with study of big data science. Crop data can be break down into bits and bytes it will give better study about the crop development by using advance data analytics tools for betterment of agriculture. Here, talk about some important tools and techniques to handle and study the big data.

Key words: Cloud Computing Analytics, Internet of Things, Big Data, Environmental Conditions, Machine Learning, Data Mining, HARVIST, WEKA System, R-software.

Big data is a collection of large and complex data sets which becomes difficult to process using on-hand database management tools or traditional data processing techniques. Big Data characterize by four components: (a) volume (how big the data is), (b) velocity (how fast the data is being collected), (c) variety (how varied data being collected is) and (d) Veracity (big data solutions must validate the accuracy of the bulky amount of rapidly incoming data). Big data requires a shift in computing architecture so that users can handle both the data storage and analyzing large volumes of data economically.

The total amount of farmland in the world is limited and according to UN estimates the global population will grow 47% to 8.9 billion in 2050. Major food grain crop yields like wheat, rice and maize are all affected in the run up to 2050 due to global warming and climate change. Research activities involving to advancement of agricultural research mainly concerning on the genomics, agricultural bioinformatics, and computational biology of plants and animals have been enabling scientists and research organizations to better feed the world populations and improve the quality of agricultural produce. These disciplines are involved in very large data sets and computationally expensive analyses.

The major problem of agricultural growth is traditional agriculture practiced which is barrier to use scientifically developed agronomic practices, such as planting, irrigating and harvesting using at suitable point in time. Weather forecasting and predictive analytics can be used by collecting real-time data on weather, soil, crop stage to make smarter decisions for crop cultivation. Advanced statistical research is able to build improved models and simulations that can predict future conditions and help farmers make proactive decisions.

Farmers need to understand how to use optimized plant density in limited cultivable

**Correspondence address : prakash289111@gmail.com*
[1] *ICAR-Indian Agricultural Statistics Research Institute*

area. Seeds sowing, fertilizing application and maintaining the crops are time-sensitive and heavily influenced by the weather. Forecasted whether determined it's going to rain or not can also influence when to irrigate fields. To reduce the food grains loss due to undetermined factors can be controlled by using suitable grain harvesting time and their transportations and distribution in appropriate time using big data analysis. Mobile app can be implemented under digital India programmed use to figure-out the precise date, time and expected amount of rainfall to avoid the seed loss due to dry or high moisture content in soil. The world population will cross over 9 billion by the year 2050, the hunger gap is expected to widen. Scientist's mission is to improve productivity through using big data analysis like weather forecasting, real-time optimization of farming machinery, monitoring of market price, automated irrigation recommendations system, mobile-based information system etc. In these fields organizations in the United States by now operate cloud-based farming information systems that apply soil observations and weather measurements to forecast weather for the next seven days. For instance, many organizations in India take part to develop a system that monitors above conditions to maintain the quality and consistency of agricultural produce.

Piatetsky-Shapiro and Frawley (1991) were described knowledge discovery in databases and defined data mining is the process of discovering previously unknown and potentially interesting patterns in large datasets. Breiman (1992) worked on meta-classifiers is a Recent developments in computational learning theory have led to methods that enhance the performance or extend the capabilities of these basic learning schemes. Boosting and bagging classifier methods used for classifications. Garner *et al.* (1995) was developed a process model for applying data mining techniques to data, with the goal of incorporating the induced domain information into a software module. Hall and Smith (1998) and the wrapper method (John and Kohavi, 1997) worked on filter algorithms provide facilities for general manipulation of attributes for example, to insert and delete attributes from the dataset. Wagstaff *et al.* (2005) were explained Heterogeneous Agricultural Research via Interactive, Scalable Technology (HARVIST) project. Their goal is to integrate multiple Earth Science data sources into a single graphical user interface that allows for the investigation of connections between different variables and focus on relationships between weather and crop yield, but the system we are creating will be capable of integrating data for other studies as well. Oswaldo *et al.* (2011) focuses on Computational solutions to large-scale data management and analysis which presents cloud and heterogeneous computing as solutions for tackling large-scale and high-dimensional data sets. Tsuchiya *et al*. (2012) introduced two fundamental technologies to analyze massive amount of data: one distributed data store and complex event processing, and other workflow description for distributed data processing. Waga and Rabah (2014) studied big data management and cloud computing analytics for sustainable agriculture. They focuses on environmental conditions' data like rainfall, winds, temperature *etc*. and the use of particular cloud computing analytical tool to get some meaningful information from it which can be utilized by farmers for strategic and successful Agriculture.

Software Tools and Techniques Used for Handle and Analysis of "Big Data"

Cloud Computing Analytics :

Cloud computing promises enough capacity in terms of storage and processing power to elastically handle data and involves delivering hosted services over the Internet. It can deals on environmental conditions data like rainfall, humidity condition, winds speed, environmental temperature etc. and the use of particular cloud computing analytical tool to get some meaningful information from it which can be utilized by farmers for strategic and successful Agriculture. The data collected ranges from soil moisture to nitrogen and other nutrients levels present in soil. The use of these data will allow farmers to gain a clearer picture of farming, receiving updates of the land in real time. For example, the data can monitor pests, which then allows farmers to target problem areas rather than over using pest

controls. To reduce water pollution cause by inadequate use of fertilizer in soil through soil analysis using cloud computing.

Illustration: Using hyper-spectral camera / aerial photography the images collected from farmer's field → Data analyzed by cloud computing service provider like Ceres Imaging (They provide farmers spectral data to optimize water and plants nutrients) → Results tell farmers how healthy their plants were, whether there was an adequate amount of fertilizer and water used or not, whether there were diseases or pests controlled or not.

Pros: Lower operational cost, less time taken, enable heavy duty data crunching to better process and better explore internet information, pay for uses to recovery fixed expenses on hardware, software development, their maintenance and support.

Cons: Security concerns, reliable internet connections, data ownership, processing responsibility, data and processing are mercy of service provider *etc.*

R

R is an open source language which is used for linear and non-linear modeling, forecasting, data mining, time series analysis, supports procedural programming with functions, matrix calculation and visualization of data. It is an interpreted language which gives results through a command-line interpreter. During the last decade R programming language has become a single most important tool for computational statistics for both the students as well as agricultural and industrial researchers. To handle the big data R have some libraries / packages like Rhdfs, Rodbc, ffbase and rmr2 *etc.* R supports many languages such as C, C++ and Python etc. to directly manipulate the R objects and it will also take less memory up to 2-4 GB only, depends on hardware configuration of system. R provides better connection with its different packages to overcome the problem of storage. To handle big data in R five strategies should be considered:

(1) *Sampling :* A good sampling strategies are sometimes very useful to reduce the size of a big data.

(2) *Requirement of bigger hardware:* By increasing the size of the machines memory. Today R can address 8 TB of RAM if it runs on 64-bit machines.

(3) *Store objects on hard disc and analyze it chunk wise:* There are different specific packages available in R which facilitates to store data in hard disc and analyzed it chunk wise.

(4) Integration of high performing programming languages such as C++ or Java *etc.* and

(5) Use of different interpreters along with R to deals big data such as pqR, TERR and R interpreter in Java etc.

Heterogeneous Agricultural Research Via Interactive, Scalable Technology (HARVIST)

The HARVIST system will provide multiple machine learning and data analysis algorithms that can be applied to the agricultural data. Here it is mainly utilized machine learning techniques which are a group of supervised / unsupervised learning methods that can be applied for classification data, clustering and dimension reduction with adding multivariate spatial modeling methods. HARVIST operates on remote sensing data which are collected through Multi-angle Imaging Spectro-Radiometer (MISR) instrument. These data are required advanced classification technique using HARVIST through support vector machine (SVM). This classifier can be used by scientists or policy makers to constructing land cover / land use (LCLU) data bases. It seeks to address both shortcomings by demonstrating the technology required to perform large-scale studies of the interactions between agriculture and climate.

As shown in Figure 1, the HARVIST system will incorporate data from weather stations, remote sensing instruments and historical crop yield databases to generate highly accurate predictions. Quickly and interactively result obtains for a large area HARVIST techniques also employ to analyze soil properties with micro-environment around crop which gives more predictive accuracy of the system.

HARVIST Analysis Methods mainly involves two data analysis methods one support vector machines (SVMs) and others clustering techni-

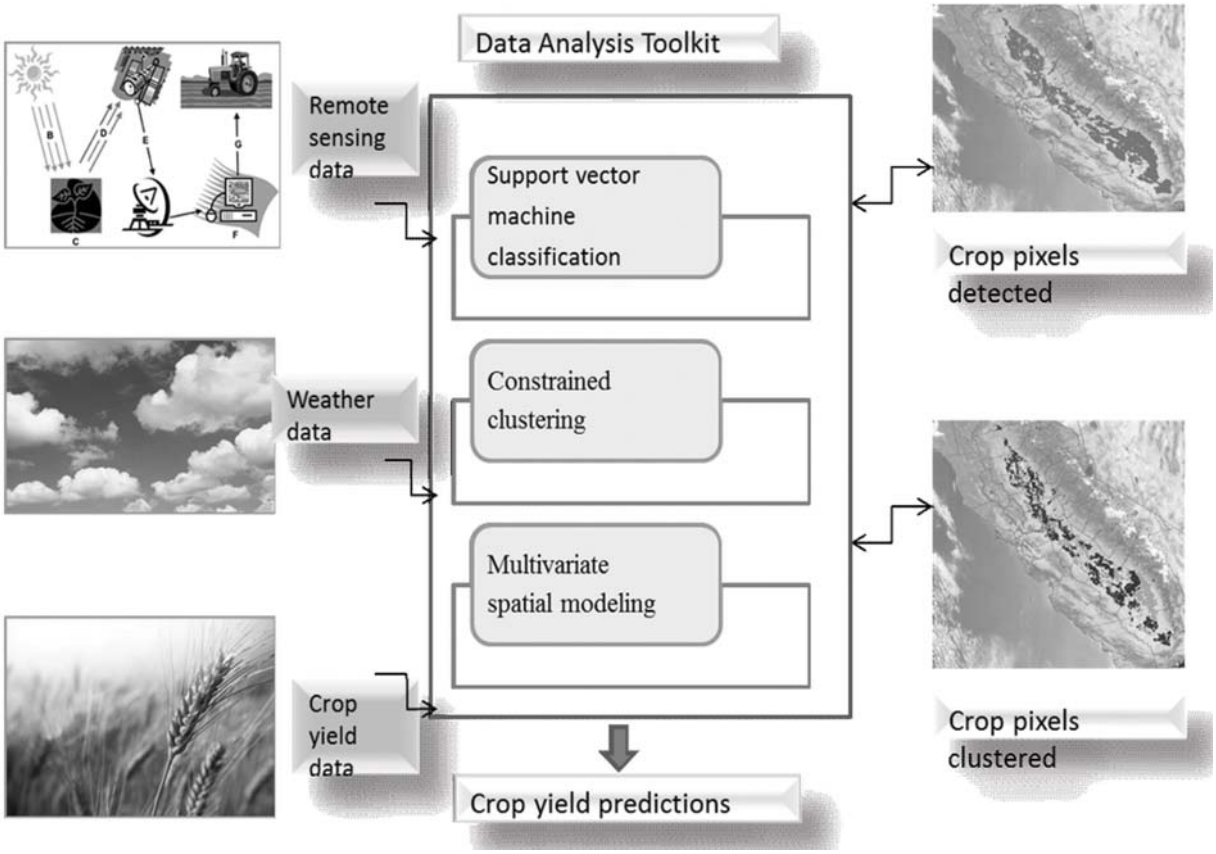

Fig. 1. The HARVIST System Architecture.

ques. These techniques are used for multispectral data from remote-sensing images. To solve pixel Classification and clustering problem: Support vector machines (SVMs) are useful for classification when the user has several specific classes of interest and the objective is to build a suitable classifier that automatically classifies big data set. In contrast, clustering methods are useful to identify overall trends present in the data set when the classes of interest are not known.

For prediction problem multivariate spatial models can be used. If samples are independent, this assumption for the given statistical models which incorporate spatial dependencies can provide more accurate predictions. The linear/ non linear model can be used to predict values for multiple response variables simultaneously and provided a straight forward method for estimating the uncertainty associated with each prediction

The Waikato Environment for Knowledge Analysis (WEKA)

The WEKA system applying data mining techniques to analyze large data sets Holmes *et al.* (1994). It can incorporate the following tools:

For data mining it is required to transform the raw data into an appropriate form by a data pre-processing routines, supporting the manipulation. Appling feature selection tools which can useful for identifying irrelevant attributes. Classifiers and some other data mining algorithms can be use to handle numerical as well as categorical learning tasks. Metaclassifiers can be used for enhancing the performance of classification. Appling bench marking tools that can be use for compare the relative performance of different learning algorithms over several datasets. To understand the WEKA system it is required to know about process of data mining.

Data Mining Process Model

Applying process model for data mining techniques to agricultural data, with the software module incorporated with induced domain information. The feature of this model can be illustrated as a two-way interaction between the data mining experts and agricultural data provider. To good fit between a representation of the data and a data mining algorithm machine learning technique has been takes several cycles through the process model. WEKA used the machine learning techniques and for understand the process we should first know the classifiers that is defined as set of rules or the form of a decision tree that can be used to predict the classification of a new data instance. ZeroR is the most primitive learning scheme which is in WEKA. For classification of new data item, it is predicts the most frequent category value in the training data for given problems with a nominal class value, or predicts the average class value for numeric prediction problems. To combining multiple models meta-classifiers can be used which is recently developing computational learning theory to enhance the performance of these basic learning schemes. Meta classifiers means to make predictions, instead of using a single classifier, it can used committee of classifiers. Two of the most prominent methods for constructing ensemble classifiers are bagging and boosting. Bagging works by building separate models of the training data using a sampling technique that deletes some instances and replicates others. But, boosting is iterative process that can use instead of sampling fresh training data, each new model is influenced by the performance of those built previously. Promotion and relegation of instances that are incorrectly and correctly classified in previous iterations correspondingly.

Clustering is technique to find the natural groupings in the dataset. Clustering is followed by a second learning stage, in which a classifier is used to induce a rule set that allocates each instance in the dataset using clustering algorithm. WEKA used the EM clustering algorithm that makes the assumption, common to other clustering algorithms, that the attributes in the dataset represent independent random variables.

Association rules in WEKA: It is a type of learning scheme commonly used in market basket analysis (MBA).This algorithms can be use to analyzing consumer purchasing patterns particularly, products detection that are frequently purchased together. These algorithms were developed in response to the large transaction data formed by barcode-based purchasing / selling systems.

CONCLUSION

Big Data bring new opportunities to modern society and plays an important role in agricultural science; producers make more informed decisions regarding crop production practices. Therefore, higher yields with lower costs can be full filled by applying the suggestion provided by crop advisors and service providers. To setup new agro-based business opportunities like industrial farming, contact farming and agro-based industries etc. can be grown by using big data analysis. It is necessary to develop the algorithms for big data analyses which can solve the problem arise in laboratory to implementing it in field. Cloud Computing analytics, R-software, HARVIST analysis methods and WEKA system are some significant tools that

Algorithm : for bagging	**Algorithm : for boosting**
• In training data: n be the number of instances	• Assign equal weight to each instances
• Perform t-time iterations for each	• Do t iterations for each instances
• Using deletion and replication for sample n instances randomly	• Model building by apply a learning technique from the weighted instances
• Model building by apply a learning technique	• Store the resultant model
• Store the model	• Down-weight all instance which correctly classified by the model
• End	• End

are used to understand how to big data use for agricultural development. The challenges to data scientists in big data analytics including storing models, information sharing, maintaining data security, data accessing, patterns finding in data, new computing to deals agricultural data, advanced analytics tools and redesigning mining algorithms *etc.* It is required to adopt emerging technologies that deals with big data challenges in agriculture.

REFERENCES

Breiman, L. 1992. Bagging predictors. *Machine Learning*, **24**: 123-140.

Fan, J., Han, F. and Liu, H. 2014.Challenges of big data analysis. *National Science Review*, **1**(2): 293-314.

Garner, S.R., Cunningham, S.J., Holmes, G., Nevill-Manning, C.G. and Witten, I.H. 1995. Applying a Machine Learning Workbench: Experience with Agricultural Databases. Proceedings of the Machine Learning in Practice Workshop, 12th *International Machine Learning Conference* (Tahoe City, CA, USA).

Hall, M.A. and Smith, L.A. 1998. Practical feature subset selection for machine learning. *Proceedings of the Australian Computer Science Conference, Perth, Australia*, 181-191.

Holmes, G., Donkin, A. and Witten, I.H. 1994. Weka: A machine learning workbench. Proc Second Australia and New Zealand Conference on Intelligent Information Systems, Brisbane, Australia.

John, G.H. and Kohavi, R. 1997. Wrappers for feature subset selection. Artificial Intelligence, **97**(1-2): 273-324.

Oswaldo T., Prins, P., Marc S. and Ritsert C.J. 2011. Big data, but are we ready? *Nature Reviews Genetics*, 12: 224.

Tsuchiya , S., Sakamoto, Y., Tsuchimoto, Y. and Lee, V. 2012. Big data processing in cloud environments. *Fujitsu Science Technology Journal*, **48**(2): 159-168.

Waga, D. and Rabah K. 2014. Environmental Conditions' Big Data Management and Cloud Computing Analytics for Sustainable Agriculture. *World Journal of Computer Application and Technology*, **2**(3): 73-81.

Wagstaff, K. L., Mazzoni, D. and Sain, S. 2005. HARVIST: A system for agricultural and weather studies using advanced statistical models. *Proceedings of the Earth-Sun System Technology Conference*. University of Maryland Inn (USA).