Fertilizer Recommendation System using SGD on Mahout and Hadoop Platform

Raghu Garg, Himanshu Aggarwal

Abstract: - In the current world scenario, the existence of human is impossible without the necessary proliferation of plants. Health of plant depends on water and soil nutrition that help plants to produce energy. Apply appropriate recommended fertilizer quantity is necessary for a healthy plant. However, due to overexposure, soil sometimes gets degraded, so fertilizer is an important element to retain the soil quality. Now a days decision support system plays a vital role in the recommendation. These recommendation systems are based on historical data. In this respect, soil analysis is an appropriate approach to determine the soil quality. Soil analysis generates a report of unstructured and unperceivable data by testing soil in laboratories that make it agriculture big data. This type of systems generally has been implemented in the banking and health care sector for fraud detection and patient recommendations, respectively. In this paper, we have been proposed fertilizer recommendation system based on present nutrition quantity in the soil. In this system, the useful data is extracted from soil analysis reports and save into two files: 1) first file save soil nutrition composition and solution number that act as the label in 2) second file save solution number and recommended fertilizer quantity. Soil composition encoded into vector use by classification system to trained system. In this research work, SGD big data analysis machine learning techniques are applied to identify the fertilizer recommendation classes based on present soil nutrition composition. Here, SGD classification system is used to train the system. Our proposed system obtained 64.08% total average accuracy. The proposed model can also be used by agriculture experts to recommend fertilizer quantity according to crop type and present nutrition composition.

Keywords: agriculture industry, big data analytics, fertilizer recommendations, Hadoop .

I. INTRODUCTION

Soil plays a vital role in plant growth, and it absorbs water and nutrition from soil that helps plants to produce energy from photochemical process. It is necessary to give appropriate fertilizer quantity to the soil to fulfil soil nutrition that helps plants to growth. To apply appropriate fertilizer quantity, first, observe present soil nutrition according to soil analysis which recommends fertilizer quantity and then applies to complete soil nutrition.

In this system, two types of recommendations are possible one is manually and second is automatically. In manual recommendation the field expert recommend the solution, accuracy of recommendation depend on his/her experience. In an automatic recommendation system, solutions are recommended by the computer machine. Automatic recommended systems are two types: a) redefined recommendation solution, b) recommend solution based on facts this type of system is called a decision support system. In this era of technology, no agriculture decision support system exists that automatically recommended fertilizer quantity based on soil nutrition composition, because this type of systems required very high computing power for analyzing historical data and large storage space to store bulk data. Hadoop framework and Mahout Machine Learning library are the solutions to the problem. Hadoop provides two type of services MapReduce [1] programming model and Hadoop distributed file system. MapReduce provides parallel job execution on a distributed environment, and HDFS provides service to store large amount of data on the distrusted network. Mahout provides machine learning library for classification, clustering and recommendation that implement on parallel Hadoop environment. This type of systems has been implemented in public sector [2], healthcare [6]and banking sector [11].

Many soil and grain classification systems have been implemented in agriculture which is based on digital image processing on small datasets shows in Table 1.

Table 1: Literature of classification system used in the past.

Author	Description
Guevara et al., 2011 [4]	They developed a machine vision system for classification of wheat and barley grain. For experiment 99 features are used where 21 features are morphological features, 72 texture features and 6 color features. This experiment conducted in two classes. This accuracy rate of this experiment is more than 99%.
Ronge et al., 2014 [13]	The ANN and K-NN classifier use to classify Indian wheat seed varieties by extracting 31 features including grey level textural features using local binary pattern, local similarity pattern, local similarity number, co-occurrence matrix etc. The average accuracy of this experiment is 66.8%.
Pazoki et al.,2013 [10]	They developed a classification system using a multi-layer perception network for classes of rain, red wheat growth with 21 statistical features. The average accuracy of the system is 87.22%.

Revised Manuscript Received on July 05, 2019

Raghu Garg, Computer Engineering/ Punjabi University, Patiala /India
Dr. Himanshu Aggarwal, Computer Engineering/ Punjabi University, Patiala /India.

Gulmezoglu et al.,2015 [5]	They use a common vector approach to classification six wheat varieties. The property of each class is computed, and a given test image is assigned to its label on minimum distance criteria. The average accurate of the system is 36.7%.
Romera et al., 2013 [12]	The classification of wheat yield using machine learning software WEKA. The experiment obtained 90% accuracy.

In this study, we found a single research paper that classification soil based on its chemicals properties [9]. Classification of soil using support vector machine (SVM) based on the known value of particular chemicals and physical properties in simple profile. Input value of this technique is the chemical composition of the soil. The advantage of this technique is to needed a small number of samples to training a system. But this technique is not implemented in soil report classification based on a solution because of these reports prepared by human experts. In this paper, we developed a fertilizer recommendation system based on present nutrition quantity in the soil. In this model, first of all, we extract features from soil analyse reports such as soil nutrition quantity and fertilizer quantity that recommended by human expert. These features are stored on the distributed system. In this model used SGD classification algorithm for classification of data. From the experiment, we found overall accuracy of this model is 64.08%.

For this experiment, we collect data from Tata soil testing laboratory. This dataset contains soil analyze reports with recommended fertilizer quantity to fulfill soil nutrition problems. In this report pH, organic carbon, phosphorous, potassium, sulfur, boron, zinc, iron, copper, magnesium soil nutrition use for analyze and recommended fertilizer composition shown in Table 2. All these reports solution prepared by human experts or solution recommended by the system that contains a redefined solution. These reports contain variation.

Table 2: Fertilizer Composition used for analysis

Urea	Nitrogen 46%
D.A.P	Nitrogen 18%, Diammonium Phosphate 46%
M.O.P	Muriate of potash K60%
Sulphur Bentonite	Bentonite Sulphur S90%
Zink Sulfate	Zink Sulfate 33%
Granubor Natur	Di-Sodium Tetra Borate Penta hydrate B14.6%

II. MATERIAL AND METHOD

The following material and methods used in our experiments.

Dataset

For experiment, we collect the data from Tata chemical laboratory. This dataset contains lakhs of soil analysis reports in text format that contain analysed soil composition and recommended fertilizer quantity. In this experiment, we select only the most commonly used recommendation with the help of agriculture expert related to wheat and paddy.

SGD

SGD stands for Stochastic Gradient Descent. SGD is a classification model for prediction of the probability of occurrence of an event. SGD can implement on Hadoop distributed parallel environment using mahout machine learning library. SGD contains three packages vector encoding, SGD learning and evolutionary package. SGD is included many other features online evolution using cross-validation, makes heavy threads use to increase machine utilisation etc.

Hadoop

An Open source project from Apache Software Foundation that has rapidly emerged as the best option to handle massive amounts of data [7,8].

MapReduce

A Programming model that enables the users to process large amounts of structured and unstructured data in parallel batches across the large cluster of the machine in reliable and fault tolerant manner [1, 15].

HDFS

HDFS is designed to store very large scale datasets on big data infrastructure on a various large cluster, thousands of servers both host directly attached storage and execute user application tasks [14].

Mahout

Apache Mahout is an open source project by the Apache Software Foundation (ASF) with the basic goal of creating scalable machine-learning algorithms that can freely be used under the Apache license. Mahout contains implementations for clustering, classification and recommendations and also for evolutionary programming [3]. Wherever necessary, it uses the Apache Hadoop library to enable Mahout to scale effectively in the cloud.

III. PROPOSED MODEL

To purpose of this research work is to develop a fertilizer recommendation system based on present nutrition quantity in the soil. This system will help a human expert to recommended fertiliser quantity based on historical evidence that fulfills soil nutrition problems. For this proposed system we used SGD classification algorithm on distributed Hadoop environment using mahout machine learning library show in figure 1.

In this model, first of all, we extract soil nutrition values as features that obtained from soil analyze and recommended fertilizer quantity using text mining

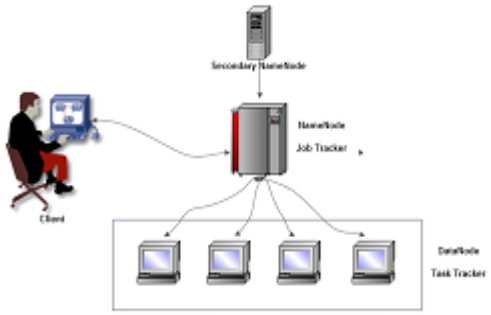

Figure1: Hadoop Cluster

technique which is in text format. Now we create two files one for soil nutrition values and second for the solution that contain recommended fertilizer quantity. In the first file, we save soil nutrition value that obtained from soil analyse and solution id that consider as the target value and save on HDFS. In the second file, we save recommended fertilizer quantity and solution id that will use for recommendation. After that
we divide the data into two parts one for training and second for testing, that will use to check the performance of the proposed model. In the recommendation process, masternode accepted the soil nutrition composition and crop name from web interface, an application program interface application and the soil testing machine. MapReduce programming model is used for parallel job processing and HDFS for distribute data storage. Job tracker distributed the jobs to data locality nodes, Mapper performed trained machine learning algorithms and predicts specific fertilizer recommendation class. Mapper used Hashmap to store the results and submitted to the reducer. Reducer shuffled the data and predicts final fertilizer recommendation class. Figure 2 shows the process of fertilizer recommendation system.

This proposed model implemented of Hadoop environment with Mahout machine learning library that will parallel job execution on a distributed environment. For data storage, we use Hadoop distrusted file system that store file on the distrusted network.

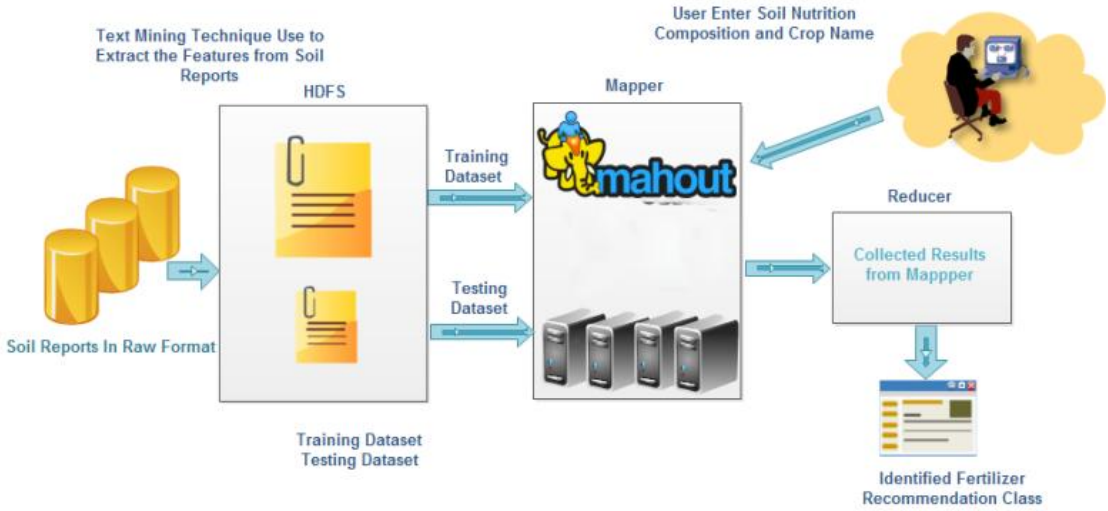

Figure 2 : Fertilizer Recommendation System

IV. EXPERIMENT AND RESULTS

For this experiment, we select the most commonly used solution for classification of soil reports with the help of agriculture expert. In this experiment, we select ten commonly recommended solutions that show in Table 3. For this experiment setup, three node Hadoop cluster with Hadoop 2.7 framework and mahout 0.7 machine learning library have been used in MapReduce programming model. It is shown in figure 1 and here eclipse IDE used for the programming environment. In this experiment, the extracted data divided into two parts: training and testing, 70% data use for training, and 30% data use for testing. In this model, thirty passes used to training model and shuffle data in each pass. From the experiment we found that overall accuracy rate of this model is 64.08% that is shown in Table 4. To obtain the accuracy measurement we use as follows:

$$Accuracy = \frac{TP + TN}{TP + FN + TN + FN} \quad (1)$$

First, we denote TP (true positive), FP (false positive), TN (true negative) and FN (false negative) as true positive (the number of instances correctly predicted as required), false positive (the number of instances incorrectly predicted as required), true negative (the number of instances correctly predicted as not required) and false negative (the number of instances incorrectly predicted as not required), respectively.

V. CONCLUSION AND FUTURE WORK

Plants must get food into their systems in order to acquire energy and continue living, similar to animals. Plants create energy for animals to use, so they must replenish their nutrients. There are many things plants need to grow such as water, nutrients, and light. Plants use water to carry moisture and nutrients back and forth between the roots and leaves. Water, as well as nutrients, is normally taken up through the roots from the soil. Fertilizer provides plants with nutrients

Table3:- Fertilizer recommendation quantity of selected classes in Kg.

Target Value		Fertilizers During Seed Sowing						Urea After Sowing	
		Urea	D.A.P	M.O.P	Sulfur Bentonite	Zink Sulfate	Granubor	After First Irrigation	After Second Irrigation
Wheat	Class I	7	56	33	11.1	0.0	2.7	29	29
	Class II	28	43	33	11.1	0.0	2.7	45	45
	Class III	2	69	33	0.0	0.0	2.7	29	29
	Class IV	12	43	33	0.0	0.0	2.1	29	29
	Class V	19	43	27	0.0	0.0	0.0	36	36
Paddy	Class VI	10	22	33	0.0	3.6	2.1	19	19
	Class VII	1	35	33	0.0	3.6	2.7	14	14
	Class VIII	9	26	33	5.8	3.6	2.1	19	19
	Class IX	4	26	33	5.8	3.6	0.0	14	14
	Class X	5	35	27	0.0	3.6	2.1	19	19

Table 4: Overall Accuracy Rate

Target Value	Accuracy
Class1	59.05
Class2	86.90
Class3	67.10
Class4	73.95
Class5	49.10
Class6	32.40
Class7	81.8
Class8	48.2
Class9	63.45
Class10	78.8
Total Average	64.08

and is usually given to plants when watering. The most important nutrients for plants growing needs are nitrogen (N), phosphorus (P), and potassium (K). Nitrogen is necessary for making green leaves, phosphorus is needed for making big flowers and strong roots, and potassium helps the plants fight off disease. In this study, fertilizer recommendation class identification system is developed to enhance the soil quality. The classification system helps in recommending accurate amounts of fertilizer to the soil based on collected soil reports that is prepared by the agricultural experts. We use SGD algorithm to the classification of soil reports, to develop this system. For performance analysis, we use overall accuracy of fertilizer recommendation system. The overall accuracy obtained from the proposed system is 64.08% discussed in table3. The accuracy rate of this system is very low cause all recommended solution is prepared by human experts with minor variations. As a future research direction, the locational attributes and more crops - other than wheat and paddy - should be integrated for a better understanding of our findings.

REFERENCES

1. A. Thusoo, A., J.S. Sarma, N. Jain, Z. Shao,, P. Chakka, S. Anthony, H. Liu, H. P. Wyckoff, and R. Murthy, R., 2009. "Hive: a warehousing solution over a map-reduce framework". *Proceedings of the VLDB Endowment*, *2*(2), pp.1626-1629.
2. EMC^2, 2014 EMC Big Data Analytics Solution For Public Sector.
3. I.Grant, 2009. "Introducing Apache Mahout: Scalable, Commercial-Friendly Machine Learning For Building Intelligen. Applications". *IBM Corporation.*
4. F. Guevara-Hernandez, J. Gomez-Gil, 2011. "A machine vision system for classification of wheat and barley grain kernels". *Spanish J. Agric*. 9 (3), pp.672–680.
5. M.B. Gulmezoglu, N. Gulmezoglu, 2015. "Classification of bread wheat varieties and their yield characters with the common vector approach." *In Proceeding of International Conference on Chemical, Environmental and Biological Sciences, Dubai*,pp.120–123.
6. A.C.Javier, C.Y.P Carmen, D.M., Robert, T. C. Stephen, T. C. Wong. and Y. Guang-Zhong, 2015. "Big Data For Health." *IEEE Journal Of Biomedical And Health Informatics*, 19(4),pp. 1193-1208.
7. S. Jianling, 2010. "Scalable RDF Store Based On Hbase And Mapreduce:. *In proceeding of 3rd International Conference On Advanced Computer Theory And Engineering. Chengdu, China*, pp. 633-636.
8. C. Lam, and J. Warren, 2010. "Hadoop in Action. Manning Publications". *Greenwich*, USA, pp. 3-5.
9. W.A. Gardner, (1984). "Learning characteristics of stochastic-gradient-descent algorithms: A general
10. Study", analysis, and critique. *Signal processing*, *6*(2), pp. 113-133.
11. A. Pazoki, Z. Pazoki, 2013. "Classification system for rain fed wheat grain cultivars using artificial neural network". Afr. J. Biotechnol. 10 (41), 8031–8038.
12. L. Phil, And G. Siemens, 2011. "Analytics In Learning And Education", *Educause Review*
13. J.R. Romero, P.F. Roncallo, P.C.Akkiraju, I. Ponzoni, V.C., Echenique, J.A., Carballido, 2013. "Using classification algorithms for predicting durum wheat yield in the province of Buenos Aires". *Comput. Electron. Agric*. 96, pp.173–179.
14. R. Ronge, M. Sardeshmukh, M., 2014. 'Indian wheat seed classification based on texture analysis using ann'. *In proceeding of International Conference on Advances in Computing. Communications and Informatics*,pp. 937– 942.
15. K. Shvachko, S. Hairong Kuang, Radia, and R. Chansler, 2010. "The Hadoop Distributed File System". *In proceeding of IEEE 26th Symposium On Mass Storage Systems And Technologies*. Incline Village, USA,pp.1-10.

16. X. Yan, Z. Wei, C. Binyue, L. Lingyun, 2015. "Research On Performance Optimization And Visualization Tool of Hadoop". *In proceeding of 10th International Conference On Computer Science & Education. Cambridge University, UK*. Pp. 149-153.

AUTHORS PROFILE

Raghu garg currently pursing Doctoral Degree in the Department of Computer Engineering at Punjabi University, Patiala, Punjab under the guidance of Dr .Himanshu Aggarwal who is serving as Professor in Department of Computer Engineering at Punjabi University, Patiala .He has done Masters in Philosophy in Computer Science from Punjabi university, Patiala, Integrated degree in Engineering from Lovely Professional University. He is a Microsoft certified web programmer and AFCH certified Ethical Hacker. He has worked on numerous live projects at during his graduation and post graduation .He is an active researcher and his current interest of research is Big data analytics and the sector chosen by him is agriculture. His published research work includes two journal and 3 research papers .His areas of interest are data mining, computer networks ,cloud computing and big data analytics..

Dr.Himanshu Aggarwal, Ph.D., is currently serving as Professor in Department of Computer Engineering at Punjabi University, Patiala. He has more than 24 years of teaching experience and served academic institutions such as Thapar Institute of Engineering & Technology, Patiala Guru Nanak Dev Engineering College, Ludhiana and Technical Teacher's Training Institute, Chandigarh. He is an active researcher who has supervised more than 35 M.Tech. Dissertations and contributed 101 articles in various Research Journals. He is guiding PhD to 8 scholars and eight have already completed their PhD degree. He is on the Editorial Board of 9 Journals and Review Boards of 5 Journals of repute. His areas of interest are Software Engineering, Computer Networks, Information Systems, ERP and Parallel Computing..