THINK INDIA JOURNAL

ISSN: 0971-1260
Vol-22-Issue-3-July-2019

Data Mining Approach For Analysis of Agricultural Research Farm Production

Barjinder Singh*
Assistant Professor
Lovely Professional University,
Phagwara, Punjab

Ravinder Singh
Assistant Professor
Lovely Professional University,
Phagwara, Punjab

Navneet Kaur
Research Scholar
Lovely Professional University,
Phagwara, Punjab

Amandeep Kaur
Research Scholar
Lovely Professional University,
Phagwara, Punjab

Abstract- **Agriculture sustainability to the ongoing change in climate and social structure is a major challenge for the researchers across the world. Agricultural system needs an Intelligent and precision agricultural approach for increasing production and productivity using limited resources. The approach needs information from various sources and efficient use of them in relevant field. The emergence of Data Mining techniques revolutionized the field of information generation and pattern recognition, it finds a wide application in agriculture and allied sectors. This research work proposed an improved algorithm in which the randomly selected cluster centroid problem is resolved using PCA. Due to this speed and efficiency of the clusters is improved.**

Keywords –**data mining, agriculture, principle component analysis, clusters.**

I. INTRODUCTION

Modern age has added large modifications and information technology in different regions of human activities have found huge software therefore also in agriculture. Development and creation of new facts technologies which allow global networking, supply agriculture the label of 'IT agriculture'. Information technology increasingly provide assistance in systematic method to solving agricultural troubles. Access to the proper data allows education of correct reviews, as an example approximately the usage of protective gadget, number of work hours of the system on a particular crop, or some of employed season work force. At the same time it's far simpler to preserve track of work and affirm change of facts. Agriculture is considerable with various facts which condition the need to apply data mining [1]. Through information mining agricultural corporations are able to producing descriptive and predictive records as guide to choice making.[2]

As is known agriculture is complex area where every day new knowledge is accumulated at increasing rate. Large portion of this knowledge is in the form of written documents, large part resulting from studies conducted on data and information acquired in agriculture from customers. Today there is a great tendency to make this information available in electronic format, converting information into knowledge, which is no easy task. With the increase in costs in agricultural enterprises and increasing necessity to control these costs, appropriate analysis of agricultural information has become the question of great importance. All agricultural enterprises need professional analysis of their agricultural data, undertaking which is long term and very expensive. Agricultural enterprises are very oriented on using information about market of agricultural products. Ability to use data in databases to extract useful

ISSN: 0971-1260
Vol-22-Issue-3-July-2019

information for high quality agriculture is a key to success of agricultural enterprises. Agricultural information systems contain massive amounts of information including information about crops, customers, market etc. With the use of data mining methods, useful patterns of information can be found in this information, which will be used for further research and report evaluation.[3] Very important question is how to classify large amount of data. Automatic classification is done on the basis of similarities present in the data. This sort of classification is only useful if the conclusion acquired is acceptable for the agronomist or the end customer. Problem of predicting production yield can be solved with data mining techniques. It should be considered that the sensor data are available for some past tense, in which appropriate production yield was recorded. All this information creates a set of data which can be used for learning ways of classifying future production yields, because new sensor data are available. There are different techniques of data mining that can be used with this purpose [5].

II. BACKGROUND

Clustering Analysis in Agriculture: This research present an effort in the direction of identifying major diseases in soybean crop, by application of cluster analysis. Data mining in agriculture is in itself a relatively new research field. The use of cluster analysis in the field of agriculture is relatively in its nascent form. It includes hierarchical agglomerative clustering approach, fuzzy clustering, hierarchical divisive clustering etc. This is a strong belief that more successful techniques can be built in the future, through intelligent use of data mining and its methods such as cluster analysis, classification and prediction, to identify the root cause factors for the spread of different diseases in different crops and also to solve various agricultural problems of different complexities and domains. [4]

PCA based Data Mining: In this research, author applied principal components Analysis (PCA) and reduced the dimensionality of the dataset and selecting the best representative subsets to improve classifier efficiency. Classification techniques in data mining are used to discover patterns and knowledge agricultural datasets, however, the accuracy of these classification techniques depends on the quality of data that are used as inputs in the data mining process. The method improves classification problems by applying principal components Analysis (PCA) and subsequently applying the Genetic Algorithm to further reduce the dimensionality of the dataset and selecting the best representative subsets to improve classifier efficiency. [6]

KNN based data mining: This research study gave an active kNNbased data mining technique is explored, also implemented to depict rural crops. Rural data extraction is a testing process, including discovering layouts, links and understanding. Data mining is relatively a new approach in the field of agriculture. Precise data in describing crops depends on climatic, geographical, natural also, other factors. These are very imperative inputs to create portrayal also, expectation models in data mining. [8]

III. PROPOSED TECHNIQUE

This section presents methodology of the proposed technique. The methodology comprises of various phases:

- Data set collection phase
- Pre-processing phase
- Classification phase
- Evaluation phase

Data set Collection: In this phase, dataset for executing proposed technique is collected. Dataset is collected from UCI repository. Dataset consists of 19 classes which represents some kind of disease. There are 35 categorical attributes, some nominal and some ordered.

Pre-processing and Filtering: Every word in a tweet is important for decision making, so effective preprocessing of these emails is an important task as these tweets are full of slang, misspellings and other language words. Therefore, in order to deal with this kind of noisy data, normalization of tweets is performed by intelligent text pre-processing techniques like tokenization, stop-word removal, stemming, lemmatization dimensionality reduction. Pre-processing of data is performed by the following techniques:

Data cleaning: Cleaning of data involves handling of missing values by ignoring that particular tuple. If any tuple or cell is empty then that will be filled with some specific value. Inconsistency of data may be handled manually. It also handles noisy data by implementing machine inspection, clustering, binning methods and regression. All the quotes (“”) from the sentences are removed, URL’s are removed and other characters that are not considered to be in the category of texts are removed. [7]

i. **Data integration:** Data is always collected from various sources like data warehouse, internet etc. so the collected data in particular so not have any use. It has to be added altogether for further analysis. So,

ISSN: 0971-1260
Vol-22-Issue-3-July-2019

this step will integrate the data collected from various sources.

ii. **Data transformation:** Transformation of data means to change the data from one form to another. For this purpose, various methods like smoothing, normalization, aggregation and generalization are available for the transformation. Transformation steps are as follows:

- Sentence Splitting: The first step involved is sentence splitting that is the splitting of string into words. Identifying sentence boundaries in a document is not a smaller task.
- Tokenization: Tokenization of words means to split a sentence into tokens or smallest unit of a sentence. Tokenization is an important task because many succeeding components need tokens clearly identified for analysis.
- Stop Word Filtering: There are a lot of words that do not have any meaning and can be removed from the input file. Words like "the", "and","for","or","if","that"; are referred to as stop words because they don't signify any meaning or sentiment. Therefore, removal of such words means stop word filtering and it also improves the performance of the system [9]
- Stemming: A stemming calculation is a procedure of semantic normalization. In this process, the variations of a word are lessened to a typical frame. For instance, consider a simple example below:

  Communication
  Communications
  Communicative----->Communicate
  Communicated
  Communicating

Classification: In the base paper, the classification can be done using KNN classification algorithm and the results are analysed. In our proposed work, the focus is on the initial cluster assignment, due to this trying to reduce the distance calculation in each iteration from all the cluster points. In the proposed algorithm, the initial clusters centres are calculated using PCA that is Principal component analysis considering the eigen values and eigen vectors calculation of each data point. The PCA algorithm is combined with the clustering technique to improve the clustering efficiency of KNN algorithm. PCA basically reduces the dimensionality of the dataset by transforming into a new set of data variables known to be as principal components, in order to extract the most relevant data information. Then enhanced KNN is applied to the dataset in order to predict the mining results. The agriculture trends can be analyzed and compared the results of the proposed technique with the existing in order to validate the results. [10]

Performance Analysis: The predicted results are analyzed and compared on the basis of accuracy and error rate.

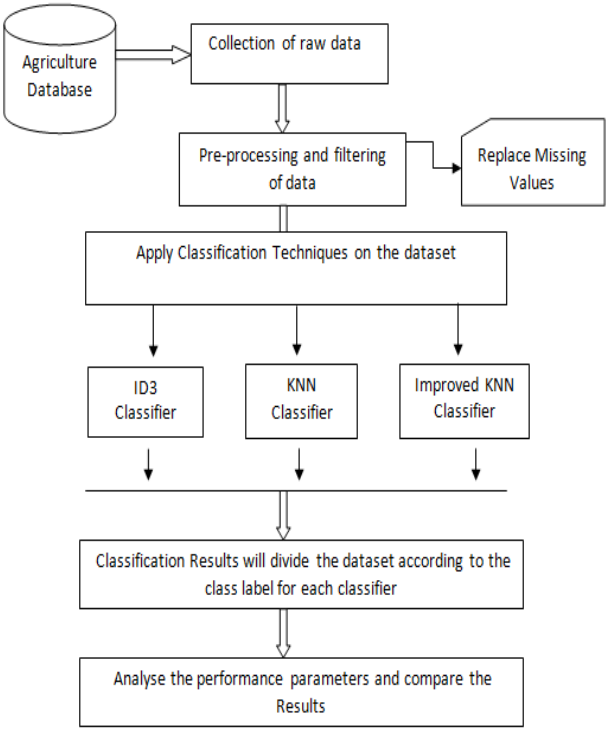

Figure 1: Flowchart of Proposed Technique

IV. PARAMETERS USED AND EXPERIMENTAL RESULTS

This section presents results of the proposed algorithm. Various parameters are used for implementation of the proposed technique. These are defined as follows:

Accuracy

Accuracy is the normal measure for arrangement execution. Precision can be estimated as effectively characterized

ISSN: 0971-1260
Vol-22-Issue-3-July-2019

examples to the aggregate number of occurrences, while mistake rate utilizes inaccurately grouped cases rather than accurately ordered cases.

$$\text{Accuracy}=\frac{\text{True Positive} + \text{True Negative}}{\text{True Positive} + \text{False Positive} + \text{True Negative} + \text{False Negative}}$$

Precision and recall

Precision and recall are mostly used to evaluate performance in text mining, and in text analysis areas like to retrieve information. These are utilized to measure exactness and completeness respectively.

$$\text{Precision}=\frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}}$$

$$\text{Recall}=\frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}}$$

F-measure

F-Measure is the harmonic mean of precision and recall. The esteem computed utilizing F-measure is stability among precision and recall.

$$\text{F measure} = \frac{2 * \text{recall} * \text{precision}}{\text{precision} + \text{recall}}$$

Mean absolute error (MAE)

MAE calculates the standard size of the error, without thinking about their bearing. This quantifies the reliability of the consta nt variables..

$$MAE = \frac{1}{n}\sum_{j=1}^{n} |y_j - \hat{y}_j|$$

Root mean squared error (RMSE)

The RMSE is a framework for quadratic scoring which determ ines the error's normal size.Communicating the recipe in words , distinguishing between figure and related watched values are all squared and the center value over the example was found af terwards. Finally, the normal's square foundation is taken.
Since the mistakes are squared prior to they are search the center value of, the RMSE provides a generally elevated weight to expansive blunders. This implies the RMSE is mainly valuable when expansive blunders are especially unwanted.

The MAE and the RMSE can be utilized jointly to analyze the variety in the mistakes in an arrangement of conjectures. The RMSE will dependably be bigger or equivalent to the MAE; the more noteworthy distinction among them, the more prominent the fluctuation in the single blunders in the example. On the off chance that the RMSE=MAE, at that point every one of the mistakes are of a similar greatness.

$$MAE = \sqrt{\frac{1}{n}\sum_{j=1}^{n} (y_j - \hat{y}_j)^2}$$

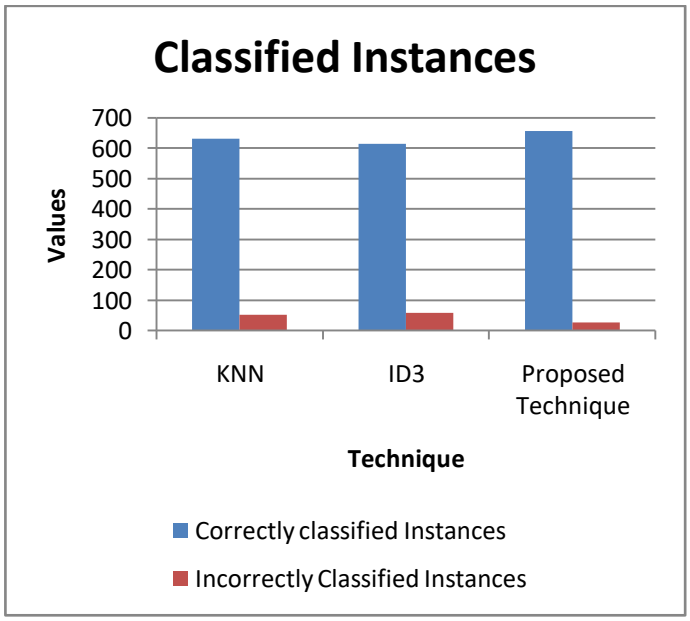

Figure 2: Showing a comparison between the proposed technique with existing technique based on classification instances

The figure above shows a correlation between the proposed technique and the current technique based on confidential instances. From the figure it is clear that the proposed technique gives more number of correctly classified instances and less incorrectly classified instances and thus better performance.

ISSN: 0971-1260
Vol-22-Issue-3-July-2019

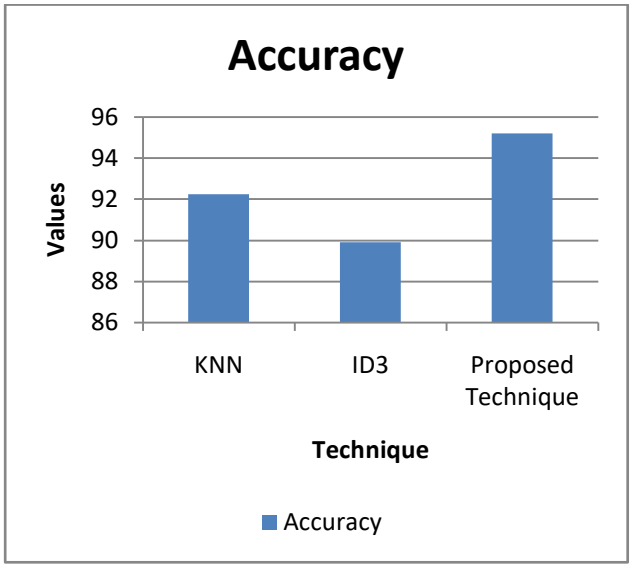

Figure3: Showing a comparison between the proposed technique and the existing accuracy technique

The figure above shows an accuracy based comparison of the proposed technique with the existing technique. From the figure it is clear that the technique proposed is more accurate and therefore performs better.

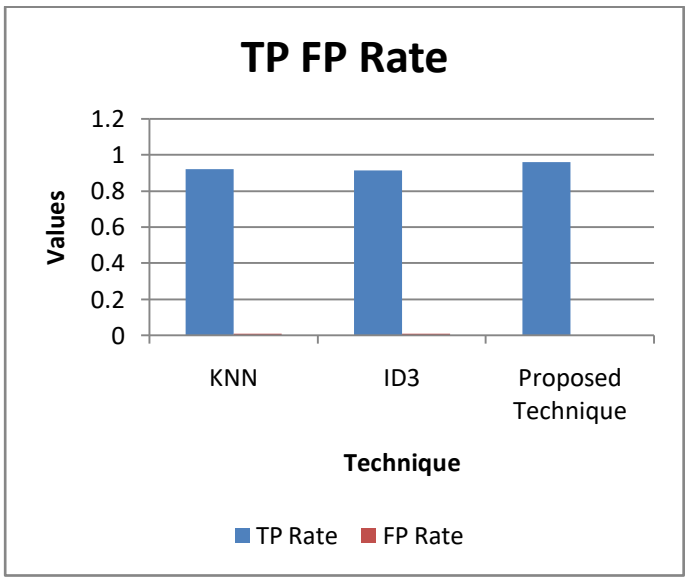

Figure 4: Showing comparison of proposed technique with existing technique based on TP rate and FP rate

The figure above indicates a correlation between the proposed technique and the current technique based on the frequency of TP and FP.

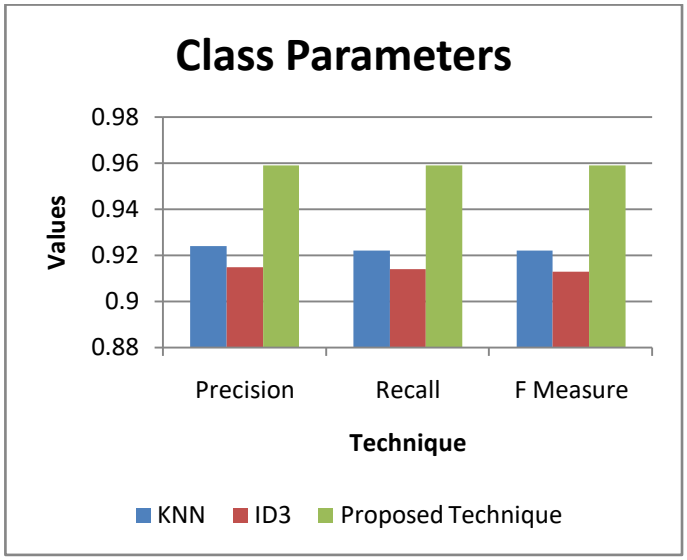

Figure5: Showing comparison of the proposed technique with the existing parameter-based technique

The figure above provides a correlation of the proposed methodology with the current parameter based technique. From the chart it is clear that the proposed technique offers more reliable recall and f-measurement and therefore better performance.

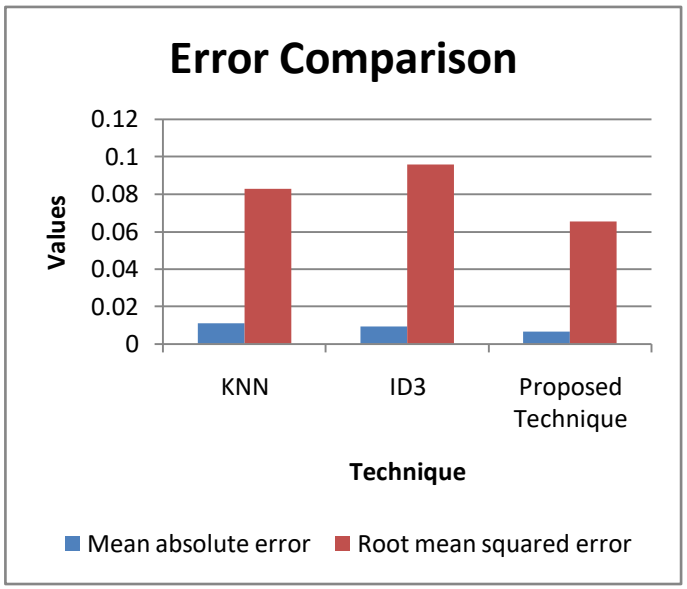

Figure 6: Showing comparison of proposed technique with existing technique based on Error

Above figure shows comparison of proposed technique with existing technique based on mean absolute error and root mean

ISSN: 0971-1260
Vol-22-Issue-3-July-2019

squared error. Proposed technique have less errors as compared to existing technique and hence performs better.

V. CONCLUSION

Data Mining is raising study discipline in Agriculture crop yield analysis. Specific information Mining procedures are in use, such as k-Nearest Neighbor (KNN), K-Means, Artificial Neural Networks (ANN) and Support Vector Machines (SVM) for extraordinarily contemporary applications of Data Mining procedures in agriculture subject. A new technique is proposed to predict the mining analysis using improved KNN classification algorithm. For this purpose agriculture data is collected and pre-processed using filtering to convert it into formatted form. Clustering is done using KNN classification algorithm. The PCA algorithm is combined with the clustering technique to improve the clustering efficiency of KNN algorithm. PCA reduces the dimensionality of dataset by transforming it into a new dataset. Then KNN is applied to the dataset to predict the mining results. The agriculture trends are analyzed and compared using the results of the proposed technique with the existing technique to validate the results. In this research work results of proposed technique is compared with KNN and ID3. Various parameters are used to evaluate the results of the proposed technique. Experimental results demonstrate that proposed technique outperforms the existing technique.

REFERENCES

[1] B. MiloviC, V. RadojeviC, "Application of Data Mining in Agriculture", Bulgarian Journal of Agricultural Science, 21 (No 1) 2015, pp. 26-34.

[2] Mona Tavakolan, Nadim W Alkharouf, "Data Mining of Genomic Data Generated from Soybean Treated with Different Phytohormones", IEEE, 2016.

[3] Savithiry S Natarajan, "Analysis of Soybean Seed Proteins Using Proteomics", Journal of Data Mining in Genomics & Proteomics, 2014.

[4] MamtaTiwari, Dr. Bharat Mishra, "Identification of diseases of Soybean Using Cluster Analysis", International Journal Of Engineering And Computer Science, 2014, pp. 5369-5372.

[5] Willyan R. Becker, Jerry A. Johann, Jonathan Richetti, Laíza C. De A. Silva, "Data Mining Techniques for Separation of Summer Crop Based on Satellite Images", Journal of the Brazilian Association of Agricultural Engineering , 2017, pp. 750-759.

[6] Dr. Geraldin B. Dela Cruz, "Comparative Study of Data Mining Classification Techniques over Soybean Disease by Implementing PCA-GA", International Journal of Engineering Research and General Science, 2015, pp. 6-11.

[7] Stacey A. Olson, Bertis B. Little, Ashley C. Lovell, "Using Data Mining to Detect Anomalous Producer Behavior: An Analysis of Soybean Production and the Federal Crop Insurance Program", pp. 1-12.

[8] P. Sharmila Devi, " Applying Data Mining Techniques to Evaluate Rural Crops Using kNN", International Journal of Advanced Research in Computer Science and Software Engineering, 2016, pp. 398-407.

[9] Jerry A. Johann, Jansle V. Rocha, Stanley R. De M. Oliveira, Luiz H. A. Rodrigues, Rubens A. C. Lamparelli, "Data Mining Techniques for Identification of Spectrally Homogeneous Areas using NDVI Temporal Profiles of Soybean Crop", 2013, pp. 511-524.

[10] Juan Yu, Zhenhai Zhang, Jiangang Wei, Yi Ling, WenyingXu, Zhen Su, "SFGD: a comprehensive platform for mining functional information from soybean transcriptome data and its use in identifying acyl-lipid metabolism pathways", 2014.