ISSN: 2277-9655
Impact Factor: 1.852

IJESRT

INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

Knowledge Discovery and Data Mining to Identify Agricultural Patterns

Kulwant Kaur[*1] , Maninderpal Singh[2]
[*1] Research Scholar, Lovely Professional University, Phagwara, Punjab, India
[2] Assistant Professor, Chandigarh University, Mohali, Punjab, India
kulwant02@live.com

Abstract

With the evolution of computer based data storage systems we have come across a huge amount of repository of data. But this data is not very helpful until we know what we can do with it. We need to make inferences from this immense data so that we can make decisions driven by knowledge. Data mining is the process of knowledge discovery in database. Mining the agricultural patterns is one of its applications. From last few decades data mining in agriculture is recent research area. Till now data mining techniques were used in the businesses and corporate sectors, but now these techniques are also being used for extraction of efficacious agricultural data. With the help of KDD and data mining we extract the meaningful data sets from the gigantic amount of data. The k-means clustering is used to classify the given set of data. This technique when applied on the large set of data then it results into improved quality of mined data. We have applied this method to study the production and consumption of crops in various parts of India. The various factors which affect the production of crops like soil type and weather are taken into consideration. For graphically representation we have used spatial join with the algorithm.

Keywords: KDD(Knowledge Discovery in Database Process), WEKA(Waikato Environment for Knowledge Analysis), CART(Classification and Regression tree), DM(Data Mining), CRM(Customer Relationship Management)

Introduction

Data mining is technique used to obtain the hidden information from huge databases. It is a powerful technique which focuses on the most crucial information in their data warehouses. According to Bhavani Thuraisingham (2009), Data mining tools anticipate the future trends and behaviors for allowing businesses to make sensible knowledge-driven decisions. The major data mining techniques are: classification, regression and clustering. Data Mining, also popularly known as Knowledge Discovery in Databases (KDD), refers to as "the nontrivial process of identifying suitable, original, useful and eventually reasonable form of data". The data is in huge amount which is stored in files, databases and other repositories. That data is progressively more important, if it is not important to build up controlling means for investigation and explanation of data and for the extraction of interesting knowledge that could help in decision-making. The following figure (Figure 1.1) shows data mining as a step in an iterative knowledge discovery process.

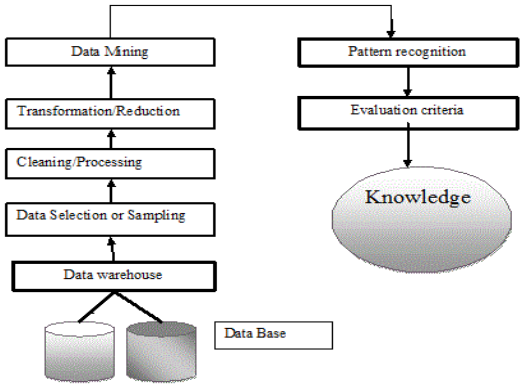


Figure 1.1 Knowledge Discovery Process

A. Knowledge discovery in data mining

According to Abdulsalam S. O., Adewole, K. S., Bashir, S.A., Jimoh, R.G. & Olagunju, M. (2012), KDD is a process which obtains the useful information from huge collection of data which make the sense of data by using appropriate techniques and methods. This process deal with the low-level data into other forms is very compact, abstract and useful.

ISSN: 2277-9655
Impact Factor: 1.852

By using the KDD process we can create short reports, modeling the process of data and predict the models that use in future.

- Data is taken from variety of sources is integrated into a single data store called target data.
- Then pre-processed the data and transformed into standard format.
- The data mining algorithms used to process the data to the output in form of patterns or rules
- Then those patterns and rules are transformed to new or useful knowledge or information.

B. Data Collection Methods

There are different techniques used in the data mining to obtain the meaningful information. We can use data mining to automatically determine significant patterns and hidden associations from huge amount of data. According to Mukesh Kumar and Arvind Kalia (2006), Data mining provides you with insights and correlations that had officially gone unrecognized or been ignored because it had not been considered possible to analyze them.

The data mining process consists of the following steps:

- **Data cleaning**: In this phase piercing data and irrelevant data are removed from the collection.
- **Data integration**: At this stage various data sources like heterogeneous data may be combined in a common source.
- **Data selection**: In this step, the relevant data is analyzed data decided on and retrieved from the data collection.
- **Data transformation**: This phase of data mining method select the data that is converted into appropriate forms for the mining process.
- **Data mining**: This is the crucial step in which clever techniques are applied to extract patterns potentially valuable.
- **Pattern evaluation**: This step is purely remarkable patterns representing knowledge are identified based on given measures.
- **Knowledge representation**: This is the final phase in which achieve knowledge is visually represented to the user. This essential step uses revelation techniques to help users understand and interpret the data mining results.

C. **Agriculture in India**

Agriculture is one of the major factors for the growth of the Indian economy. The Indian economy is mainly based on the agriculture. The annual production of products from the agricultural sector is an important factor in the growth of the economy. So without agriculture and agriculture-based products, the Indian economy cannot sustain its growth rate. The production of food grains are increased in India as the population increased. The growth rate of food grains increased due to use of fertilization and pesticides. In this research we focus our attention on production and consumption of main crops in India. We discuss the growth rate of agricultural crops and factors which affect the growth rate of major crops.

Area under major crops in India

Crops	Area
Wheat	**266 lakh hectares**
Rice	**391.17 lakh hectares**
Cotton	**83.73 lakh hectares**
Maize	**63.7 lakh hectares**
Bajra	**91.81 lakh hectares**
Sugarcane	**43.54 lakh hectares**

Figure 1.2 Major Crops in India

Wheat

Wheat and rice production and consumption rate in India. Wheat and rice are the major crops in India. The main wheat growing Zones are: NHZ, NWPZ, CZ, PZ, SHZ, and NEPZ. Main wheat producing states are Punjab, Haryana, Uttar Pradesh and Rajasthan, which are located in the northwestern zone of India. Ira Matuschke and Matin Qaim (2006), other states which are denoted as large wheat areas are: Gujarat and Madhya Pradesh in the central zone, Bihar in the eastern zone and Maharashtra in the peninsular zone.

Cotton

The cotton grows in those areas in which rainfall is less. The main 6 states grow cotton include Gujarat, Maharashtra, Rajasthan, Punjab, Haryana and Karnataka. Gujarat is number one cotton producing Indian state due to lack of rainfall.

D. Data Mining In Agriculture

Data mining is the process of identifying the previous unknown and interesting patterns in vast datasets. The required and useful information is used for representing the model for purpose of prediction and classification. Data mining is mainly categorized as descriptive and predictive data mining. In the agriculture area mainly predictive data mining is used. There are two main techniques namely classification and clustering. Data mining technique is used for the prediction of future production of crops.

Data Mining Algorithms

To create the data mining model, a data mining algorithm first analyze the data then creates the particular types of patterns .the data mining model use this result to analyze and define the

ISSN: 2277-9655
Impact Factor: 1.852

optimal parameters. These parameters are used to extract the meaning full information. The mining model that an algorithm creates from your data can take various forms.

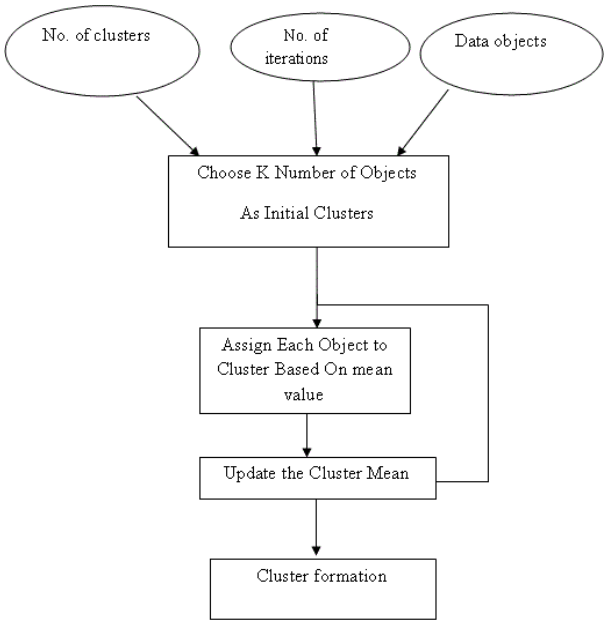
- A set of clusters that illustrate how the belongings in a dataset are related.
- A decision tree that forecast the result, and explain how dissimilar criteria affect that result.
- A mathematical model that predict the result.
- A set of regulations that describe how goods are grouped together in a transaction, and the probabilities that products are purchased together.
  - **Classification algorithms** forecast one or more distinct variables, based on the other attributes in the dataset.
  - **Regression algorithms** forecast one or more constant variables, such as profit or loss, based on other attributes in the dataset.
  - **Segmentation algorithms** partition data into groups, or clusters, of items that have related properties.
  - **Association algorithms** discover correlations between dissimilar attributes in the dataset. The main application of algorithm is creating association rules and used in a market basket analysis.
  - **Sequence analysis algorithms** review the numerous sequences in data, such as a Web path flow.

E. K-means algorithm

The *k*-Means algorithm is distance-based clustering algorithm which separate the data into a prearranged number of clusters (provided there are enough distinct cases).

Distance-based algorithms rely on a distance metric (function) to measure the similarity between data points. The distance metric is measured as either Euclidean or Fast Cosine distance. Data sets are assigned to the nearest cluster according to the distance metric used. The algorithm randomly selects k points as the initial cluster for centers.

1. Each point in the dataset is assigned to the nearest cluster which is based upon the Euclidean distance between the each point of data and each cluster center.
2. Then cluster center is recomputed by the find the average of the points in that cluster.
3. Repeat the steps 2 and 3 until the clusters converge. Convergence of clusters may be defined in a different way depending upon the implementation.



4. It is normally mean that no observations can change the clusters when steps 2 and 3 are repeated.

Literature Survey

A. Banumathi, A. Petalakshmi (2012) has explained how data mining can be applied on large datasets to discover patterns using the method of clustering. They have analyzed the Fuzzy C-Means algorithm and have made inferences that initial seed value selected either sequentially or randomly have effect on the value of the ensuing cluster.

Amrender kumar (2004) explains the techniques for forecasting of crops. The data mining application agriculture use more than one application for implementing the results. The various data mining and neural network techniques are used for forecasting the result. These techniques are decision tree, rule induction, navie-bayes, neural network, ANN, radial basis function, recurrent network, multilayer perceptron and RBF techniques and predict the result.

Anwiti Jain, Anad Rajavat, Rupali Bhartiya (2012) has explained about the clustering mechanism which is an unsupervised technique of learning. The clustering enables to identify groups based on attributes values. The groups are uniform in terms of objects they contain in them. In their research they have pointed out the use of K means clustering algorithm with modification to find the cluster's centre. They have applied the algorithm to very large data sets. They have shown that clustering has many ways of application in data mining except the way they have used. Use of iterative clustering mechanism has influence of the cluster centre chosen

ISSN: 2277-9655
Impact Factor: 1.852

for each iteration. They have used the approach of optimization formulation of problem in designing the algorithm together with novel iterative method. The research paper shows improvement in cluster centre detection when tested on large random datasets.

Dr. Rajesh (2011) explains the application of data mining in agriculture. Data mining is the great technique currently used in agriculture and industries. In this research the researcher explain the k-means clustering to classify the patterns. In this research they discuss the particular area and analysis on this areas agriculture patterns and obtain the required result. Association technique is used for the clustering.

Research Methodology

There are different kinds of parameter used for the collection, classification, regression and clustering in data mining. K- Means is the unsupervised clustering algorithm. It is simple way to apply the clustering on the different data sets to obtain the number of clusters. The result of the clusters depends on the number of data sets. The different number of data sets obtains the different result. Cluster the different data sets to its nearest center. If the clusters are far away from the center then again calculate the centroid. For identification of agricultural patterns some of these methods are used as: decision tree, association rule, clustering algorithm like fuzzy logic clustering, k-means clustering and hierarchal clustering algorithms. To identify the production and consumption of rice and wheat we use the K-means clustering algorithm to cluster the data from last few years. First we collect the data and create the database and access that data by using the java. Then apply the k-means clustering algorithm. Code that algorithm in java and run the java file into the software tool WEKA.

Proposed Algorithm

Step 1: Randomly choose a set of initial centres C0 = {C1,C2,...,Ck} and produce a set of initial weights W0 = [w0 1,w0 2,...,w0 m] (m j=1 wj =1).
Step 2: Compute initial partition matrix Set t =0;
Step 3: Let ˆ Z = Zt and ˆ W = Wt. Solve Problem P (U, ˆ Z, ˆ W) to obtain partition matrix Ut+1.
If P (Ut+1, ˆ Z, ˆ W) = P (Ut, ˆ Z, ˆ W), output (Ut, ˆ Z, ˆ W) and stop;
Else
Go to Step 3;
Repeat steps 3 to 4 for adjusted β values:
Step 4: Letˆ U = Ut+1 and ˆ W = Wt. Solve Problem P (ˆU, Z, ˆ W) to obtain Zt+1
If P (ˆU, Zt+1, ˆ W) =P (ˆU, Zt, ˆ W), output (ˆU,Zt, ˆ W) and stop;
Else
Go to Step 4;
Step 5: Letˆ U = Ut+1 and ˆ Z = Zt+1. Solve Problem P (ˆ U, ˆZ,W) to obtain Wt+1
If P (ˆ U, ˆZ, Wt+1) =P (ˆ U, ˆZ, Wt), output (ˆ U, ˆZ, Wt) and stop;
Else
Set t = t +1and go to Step 2.

Results

In the proposed work WEKA and JAVA is used to produce the required result. WEKA is the open source and command line interface. WEKA uses the machine learning techniques to classify the different data set. WEKA provides the graphical user interface and the java code to provide the clustering result. Code the k-means algorithm in java and the save the java file in .jar file then open the WEKA and run the code and provide the clustering result. We use the eclipse to code the java.

First we use the java to code the algorithm and produce the result. We use the tool WEKA to run a code. WEKA is the open source used for the purpose of coding. In java we create the frames and objects to visualize the clusters. To produce the output we compare the two algorithms and obtain the result which compare the time taken by both of algorithms to produce the output.

The figure1.3 shows the ten years data related to wheat, cotton and rice. The growth rate of production is shown in the graph. There production and consumption of the years are changes depends upon the number of factors like rainfall, use of fertilization.

Figure 1.4 shows rice and wheat production, yield and area. We extract the major crops whose production is more than other crop. Extract the data using the inner joins from the database. Rice and wheat are the main crops so we extract these main crops and produce the desired result.

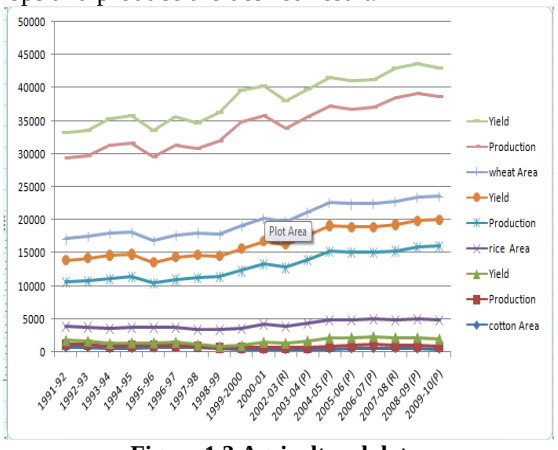


Figure 1.3 Agricultural data

ISSN: 2277-9655
Impact Factor: 1.852

Figure1.5 shows clustering result for year wise rice yield. In this x axis shows the different years and y axis shows the rice yield in this different ten years. In this we have taken area wheat, year wheat and production wheat and yield wheat. We have taken three types of temperature of wheat. There are different types of soils but in this we have taken only one type that is alluvial type of soil to generate this result.

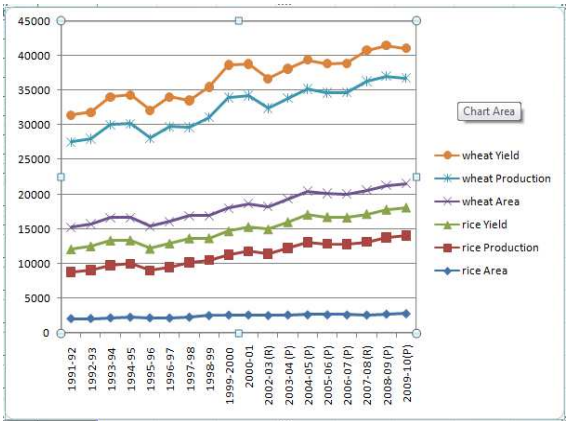


Figure 1.4 Rice and Wheat Data

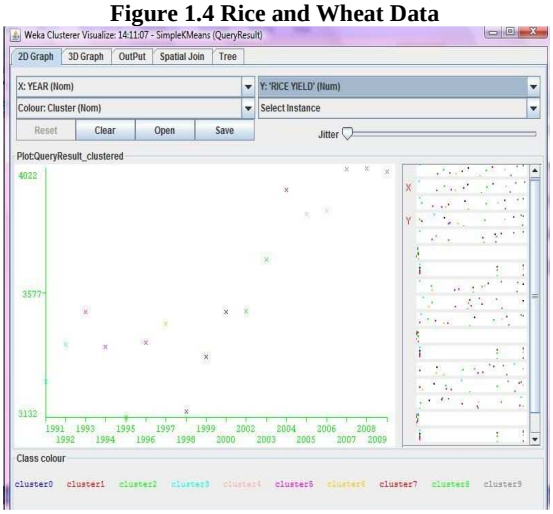


Figure 1.5 Years and Rice Yield

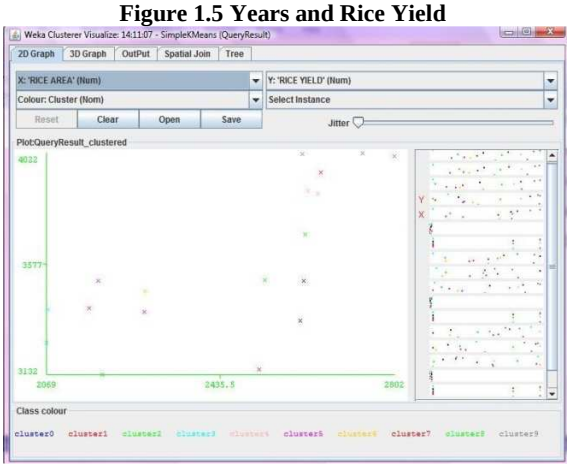


Figure 1.6 Rice Area and Rice Yield

The above figure1.6 shows the clustering results of rice yield and rice area. The x axis shows the rice area and y axis shows rice yield. This dialogue box contains the results of total production of rice. In this dialogue box on x-axis there is 'Rice Area' and on y axis we have taken 'Rice Yield'. This clusters show the total production of the rice.

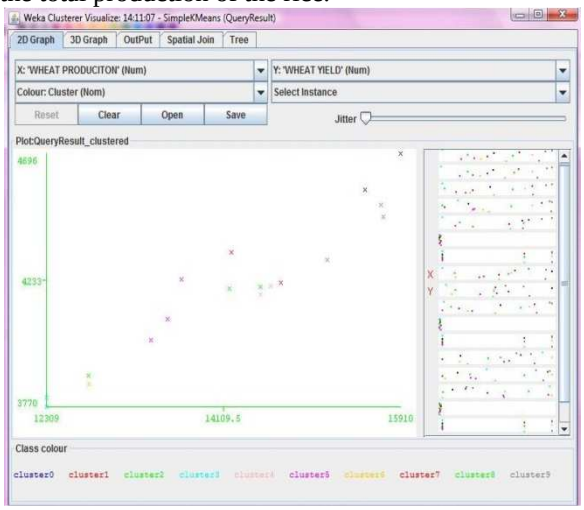


Figure 1.7 Wheat Production and Wheat Yield

This result shows the wheat production and wheat yield in previous years. In this dialogue box there are numerous clusters that is cluster 0 , cluster 1, cluster 2, cluster 3, cluster 4, cluster 5 , cluster 6 , cluster 7, cluster 8, cluster 9.These clusters show the production of the wheat of the year. These clusters are in different colors. In this if the production is less may be it will increase in future. Because sometimes we use different types of soils, fertilizers, quality of soil. These will affect the production of the wheat. Sometimes the weather changes effect the production also.

The figure 1.8 shows the three dimension clustering. In this diagram x axis shows the rice area, y axis shows the wheat area and z axis shows the wheat production.

ISSN: 2277-9655
Impact Factor: 1.852

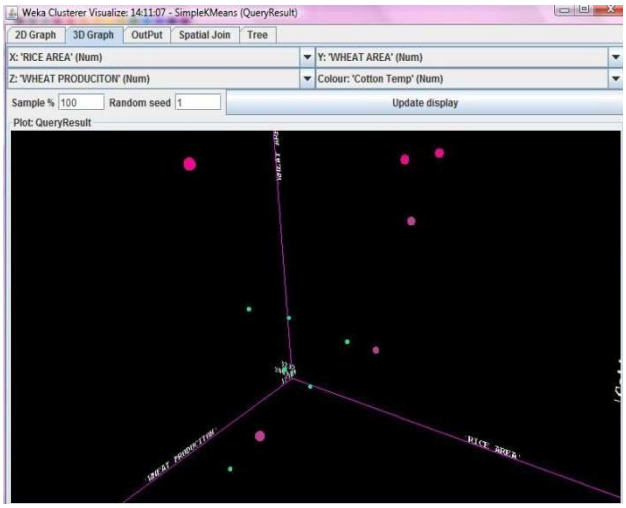


Figure 1.8: Rice Area, Wheat Production and Wheat area.

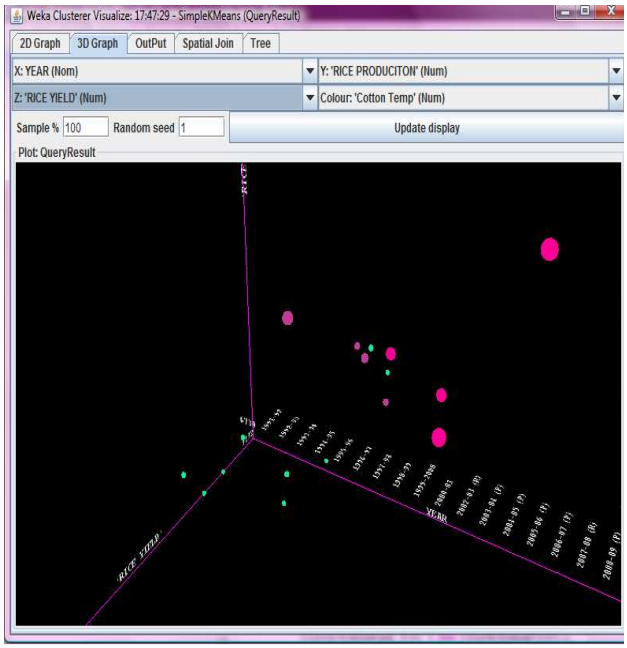


Figure1.9: Year, Rice Production and Rice Yield

This clustering result shows the year, rice production and rice yield. These three factors show the results of the rice production and rice yield in a year. These different clusters show the different results.

In the figure1.10 x axis shows the rice soil type, y axis shows the rice production and z axis shows the rice area. In this dialogue box we have three factors like 'Rice Area', 'Soil type', 'Rice Production'. These three factors show the results

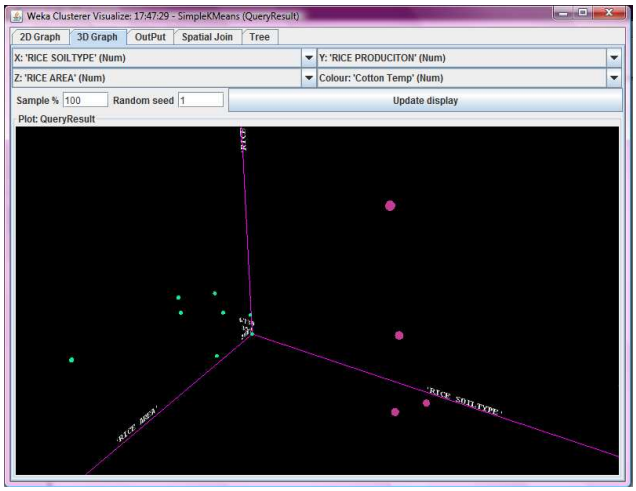


Figure 1.10 Rice Soil Type, Rice Production and Rice Area

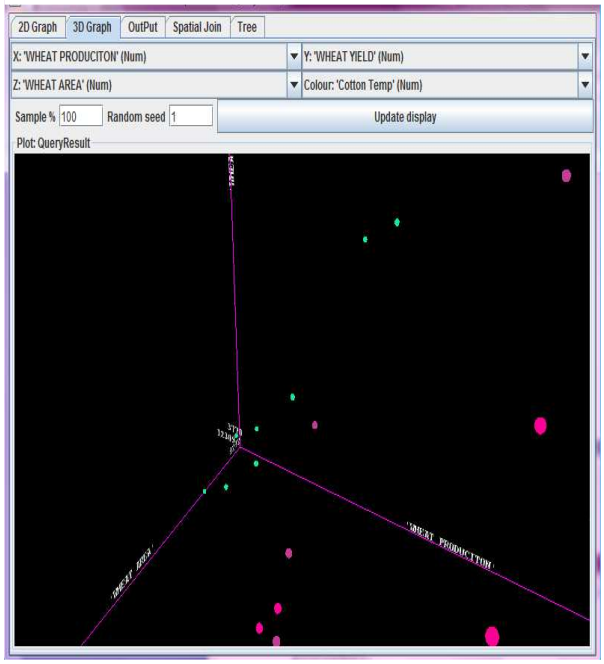


Figure 1.11 Wheat Production, Area and Yield

This figure1.11 shows the clustering of wheat production, wheat area and wheat yield in different years. In this dialogue box we have three factors like 'Rice Area', 'Soil type', 'Rice Production'. These three factors show the results.

ISSN: 2277-9655
Impact Factor: 1.852

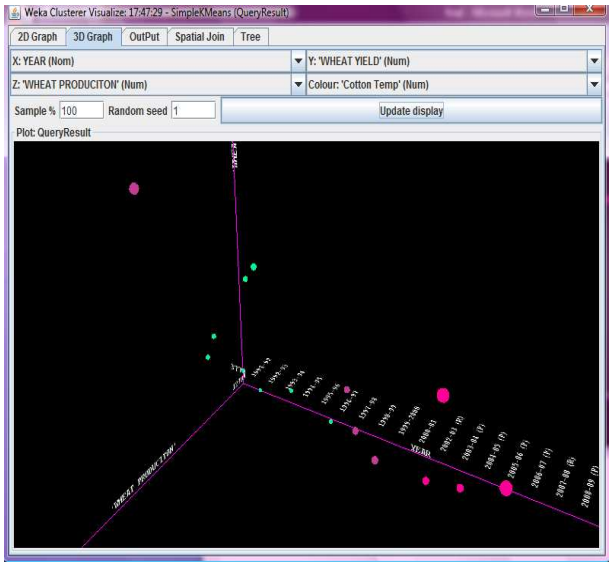


Figure 1.12 Year, Wheat Yield and Production

In the above result window clustering is shown on the basis of years, wheat yield and wheat production. In this dialogue box we have three factors like 'Rice Area', 'Soil type', 'Rice Production'. These three factors show the results.

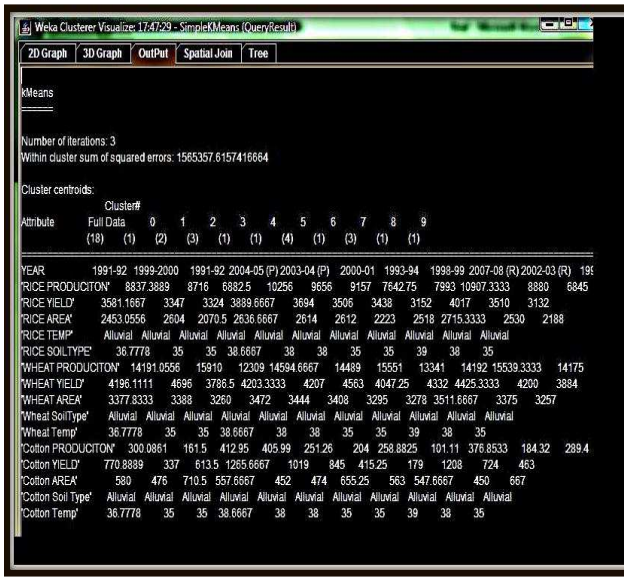


Figure 1.13 Output of Clusters.

The above result window shows the output of the clusters. This will shows the number of iterations and sum of squared error result. Also shows the result of cluster centroids and clusters.

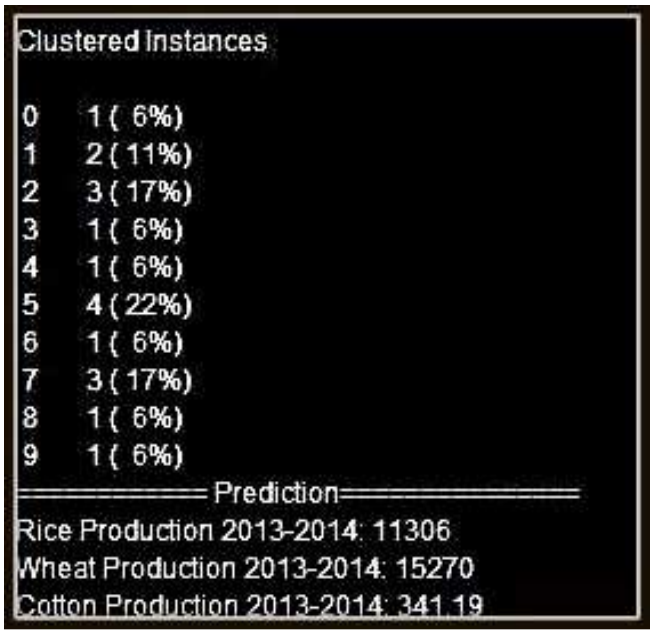


Figure 1.14 Cluster Instance and Prediction

The above diagram shows the various cluster instances and prediction of the rice, wheat and cotton production.

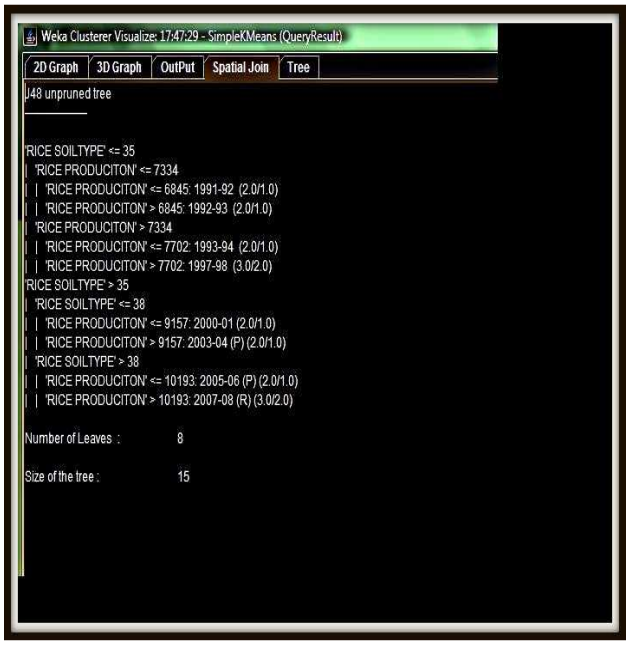


Figure 1.15 Output of Spatial Join

The above window shows the result of the spatial joins. The spatial joins applied on the dependencies on the production into the particular area. This will shows the results into the form of tree. In this result is obtained on the basis of rice production. After calculating the dependencies the size of tree is fifteen and number of leaves in tree is eight.

ISSN: 2277-9655
Impact Factor: 1.852

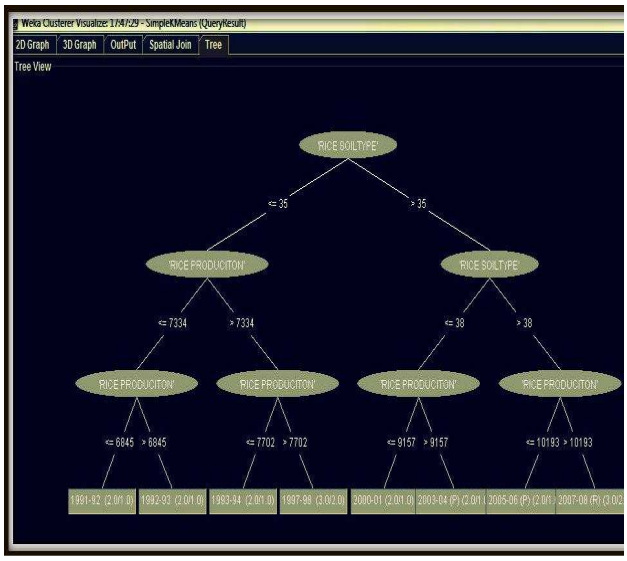


Figure 1.16 Spatial Join in Tree

The above result window shows the view of the tree on the basis of soil type and production. In this tree the left side is Rice Production and the Right side Rice soil type. In this tree the size of the tree is fifteen and the numbers of leaves are eight.

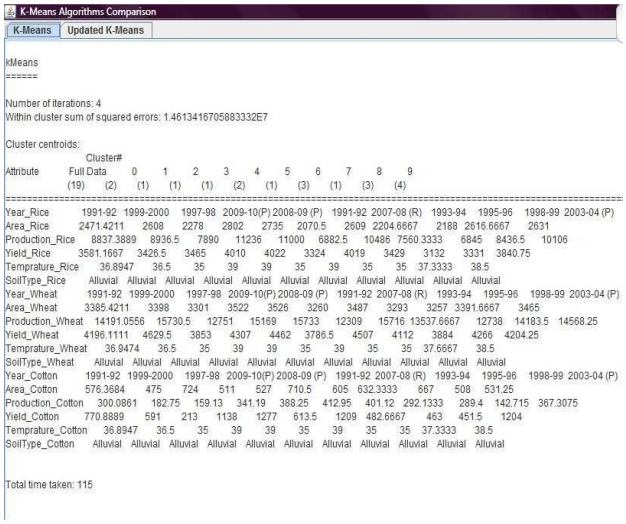


Figure 1.17 k-means clustering data and result

This result shows the time taken by the k-means algorithms to calculate the clusters. In the above figure shows the number of iterations and cluster squared errors. Then cluster centroids and attributes.

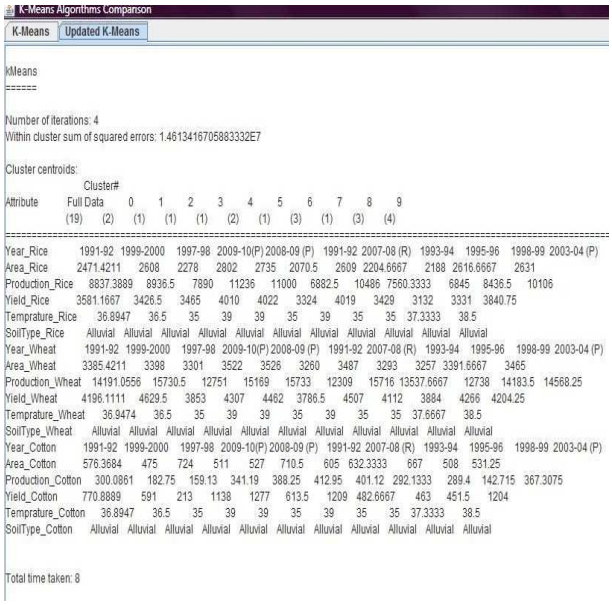


Figure 1.18 weighted k-means result

This result shows the time taken by the weighted k-means to calculate the cluster. The time is shown in millisecond. In the above figure shows the number of iterations and cluster squared errors. Then cluster centroids and attributes.

Conclusion

The work focus on identifying the required patterns from large datasets in agriculture and also finding the small patterns for identifies the production of crops. It contains the basic functionality of KDD and data mining. This work includes the techniques which, follow to extract the information, analysis on that information, follow the mining process for the identification of patterns. In this we used the data mining tools and algorithms to mine the data and also applied the dependencies on various factors. In the future work we can add some spatial data and algorithms to identify the required patterns and compare the previous algorithms and design the tree like R-tree, R* tree to how the dependencies. We can also use this technique in medical, university management system, marketing etc.

Refrences

[1] Abdulsalam S. O., Adewole, K. S., Bashir, S.A., Jimoh, R.G. and Olagunju, M (2012), Data Mining Techniques for Knowledge Discovery From Financial Institution Database.

[2] A.Banumathi, A. Pethalakshmi, (2012) "Increasing Cluster Uniqueness in Fuzzy C-Means through Affinity Measure",

Proceedings of the International Conference on Pattern Recognition, Informatics and Medical Engineering, March 21-23, 2012.

[3] Anwiti Jain, Anand Rajavat, Rupali Bhartiya, (2012) "Design, Analysis and Implementation of Modified K-Mean Algorithm for Large Data-set to Increase Scalability and Efficiency," 2012 Fourth International Conference on Computational Intelligence and Communication Networks.

[4] Bhavani Thuraisingham (2000), "A Prime for Understanding and Applying Data Mining", IT Professional Volume 2 Issue 1, January.

[5] Dr. D. Ashok Kumar, N. Kannathasan,(2011) "A Survey on Data Mining and Pattern Recognition Techniques for Soil Data Mining", IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 3, No. 1, May 2011.

[6] Dr. Mohammad A. AL-Hamami, Dr. Soukaena H. Hashem, (2009) "Applying Packets Meta Data for Web Usage Mining".

[7] Ira Matuschke and Matin Qaim, (2006), "Adoption and Impact of Hybrid Wheat in India", International Association of Agricultural Economists Conference, Gold Coast, Australia, August 12-18, 2006.

[8] K. Kamaraj, C. Chandrasekar, (2011) "Segmentation Based Sequential Pattern, Mining in Temporal Databases", European Journal of Scientific Research.

[9] K.S.Deepak, K.Gokul, R.Hinduja, S.Rajkumar, (2013) "An Efficient Approach to Predict Tumor in 2d Brain Image Using Classification Techniques".

[10] M. Chitralegha, Dr. K. Thangavel, (2013) "Protein Sequence Motif Patterns using Adaptive Fuzzy C-Means Granular Computing Model", "Proceedings of the 2013 International Conference on Pattern Recognition, Informatics and Mobile Engineering", February .

[11] M. Mayilvahanan, M. Sabitha, (2013) "Estimating the availability of Sunshine Using Data Mining Techniques", 2013 International Conference on Computer Communication and Informatics (*ICCCI* - 2013), Jan. 04 – 06, 2013.

[12] Mukesh Kumar and Arvind Kalia (2011), "Mining of Emerging Pattern: Discovering Frequent Itemsets in A Stock Data" International Journal of Computer Technology and Applications.

[13] P. Revathi and Dr. M. Hemalatha, (2011) "Efficient Classification Mining Approach for Agriculture" International Journal of Research and Review of Information Sciences Vol. 1 No. 2, June 2011".

[14] Qiang Wang, Yunming Ye, Joshua Zhexue Huang, "Fuzzy K-Means with Variable Weighting in High Dimensional Data Analysis", The Ninth International Conference on Web-Age Information Management.

[15] Vibha Maduskar, Prof. Yashovardhan Kelkar, (2012), "Survey on Data Mining", International Journal of Emerging Technology and Advanced Engineering, February.

Web Links

http://www.indiastat.com/agriculture/2/stats.aspx
http://www.indexmundi.com/agriculture/?country=in&graph=production
www.mapsofindia.com
http://www.spectrumcommodities.com/education/commodity/statistics/world%20wheat/indiawtable.html
http://faostat3.fao.org/faostat-gateway/go/to/home/E