Review Of Research
ISSN 2249-894X

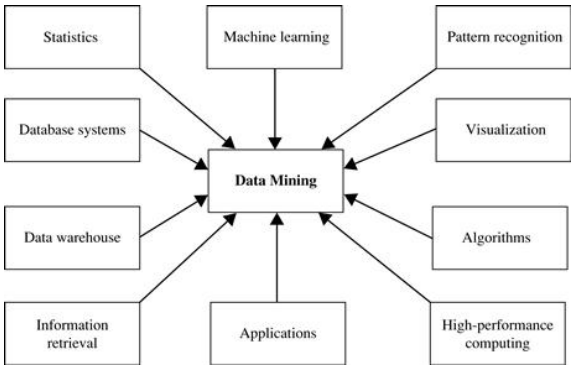
Impact Factor : 3.1402(UIF)
Volume - 5 | Issue - 1 | Oct - 2015

IN DIFFERENT PHASES OF CROP CULTIVATION USE OF DATA MINING ALGORITHMS

Md. Atheeq Sultan Ghori

Assistant Professor , Department of CSE , Telangana University , NIZAMABAD.

INTRODUCTION



Data mining can be defined as the process of selecting, exploring and modeling large amounts of data to uncover previously unknown patterns. In the agriculture sector, data mining can help farmers to gain profit and country development. For example, by applying data mining techniques, government can fully exploit data about farmers' buying patterns and behavior – as well as gaining a greater understanding of their land to protect them managing invertebrate pests and vertebrate pests, diseases, improve underwriting and enhance risk on crop cultivation. This paper discusses how farmers can benefit by using modern data mining methodologies and thereby reduce costs, increase profits, acquire new farmers, retain current farmers and cultivate new crops. Data mining methodology often can improve upon traditional statistical approaches to solving business solutions. For example, linear regression may be used to solve a problem because insurance industry regulators require easily interpretable models and model parameters. Data mining often can improve existing models by finding additional, important variables, identifying interaction terms and detecting nonlinear relationships.

Models that predict relationships and behaviors more accurately lead to greater profits and reduced costs. Specifically, data mining can help agriculture firms in production practices such as:

- Acquire new farmers.
- Retain current farmers.
- Performing sophisticated classification.
- Correlation between crops scheme.

ACQUIRING NEW FARMERS

An important business problem is the acquisition of new farmers. Although traditional approaches involve attempts to increase the farmers base by simply expanding the scheme by government, crop pattern that are guided by more quantitative data mining approaches can lead to more focused and more successful results. A traditional cultivation scheme is to increase the number of farmers by simply targeting those who meet certain scheme constraints. A drawback to this approach is that much of the human resource effort may yield little return.

At some point, average yield become more difficult and greater financial budgets lead to lower and lower returns. Hence in this situation it is important to identify population segments among already insured farmers through which uninsured farmers could be targeted. A statistical technique called “cluster analysis,” sometimes used in the private sector to identify various market segments, was used to identify target groups of uninsured adults based on the previous available data of scheme holders. Clustering is a technique of partitioning or segmenting the data into groups that might or might not be disjointed. The clustering usually accomplished by determining the similarity among the data on predefined attributes. The most similar data are grouped into clusters. Since clusters are not predefined, a domain expert is often required to interpret the meaning of the created clusters.

Government can group its crop variety based on common features. Government does not have any predefined for this label. Based on the outcome of the grouping they will target production and average yield campaigns to the different groups for a particular type of scheme.

The information they have about the farmers include survey number, crop name and variety. Depending on the type of crop scheme, not all attributes are important. For example, suppose cultivation on cotton, we could target the farmers having less water resource and human resource. Hence the first group of farmers is having constant supply of fresh water for irrigation, soil of low fertility and means temperature 70°F is suitable for Paddy. The second group is mixture of clay, mean annual temperature of over 60°F is for Cotton. Others are pulses.

DEFINITION

Given a database D = {t1,t2,…,tn } of tuples and an integer value k, the clustering problem is to define a mapping f : D → {1,…,k}where each ti is assigned to one cluster kj, 1≤ j ≤ k.A cluster kj, contains precisely those tuples mapped to it that is,
kj= { ti | f(ti) = kj, 1≤ i ≤ n, and ti D}

k-means Clustering Algorithm

K-means is an iterative clustering algorithm in which items are moved among sets of clusters until the desired set is reached.

Input:
D= {t1, t2, t3,...,tn} //Set of elements k //Number of desired clusters

Output:
K //set of clusters
Algorithm:
assign initial values for means m1, m2, …, mk;
repeat
assign each item ti to the cluster which has closest mean;
calculate new mean for each cluster;
until convergence criteria is met.

The data has been preserved from records in Agriculture Department, Perambalur. The collected data has been entered and analyzed using weka(machine learning).

III. Retaining the farmers

As acquisition rainfall decrease, crop cultivation is beginning to place a greater emphasis on farmer retention programs. Experience shows that a farmer have greater than average yield on holding two or more crops yield is much more likely to rewards than is a farmer holding an average yield. By offering quantity crops to farmers, adds value and thereby production increases farmer flexibility, reducing the likelihood the farmer will not switch land for sale to building promoters. So we have determined the frequent item sets based on a predefined support. We have all the riders that are often distributed. We need to find all the associations where farmers who bought a subset of a frequent item set, most of the time also bought the remaining items in the same frequent item set. Association refers to the data mining task of uncovering relationships among data. Data association can be identified through an association rule.

Association rule mining problem is defined as follows:

D ={ t1,t2,…,tn } is a database of transactions. Each transaction consists of I , where { i1,i2,… in}= I is a set of all items. An association rule is an implication of the form A⇒B , where An and B are itemsets, A⊆I ,B⊆I ,A∩B=Ø .In help certainty structure, every affiliation manage has support and certainty to affirm the legitimacy of the run the show. The help indicates the event rate of an itemset in D, and the certainty signifies the extent of information things containing B in all things containing An in D.

Sup(i) = Count(i) / Count(DBT)
Sup(A ⇒ B) = Sup(A∪ B)
Conf (A ⇒ B) = Sup(A∪ B) / Sup(A)

At the point when the help and certainty are more prominent than or equivalent to the pre-characterized limit Supmin and Confmin, the association rule is considered to be a valid rule. The objective of ARM is to find the universal set S of all valid association rules.

Apriori Algorithm The Apriori algorithm is the most well-known association rule algorithm and is used in most commercial products. Input:

Li-1 //Large itemsets of size i – 1
Output:
Ci //farmers of size i
Algorithm:
Ci = ø;
for each I _ Li-1 do
for each J _ _ Li-1 do
if i – 2 of the elements in I and J are equal then
Ck = Ck U {I U J};
TABLE 1. Best rules found in apriori

Crop name	Crop variety name	Support	Confidence
Cotton	K10	64	1
Paddy	Adt45	62	1
Groundnut	Tmv7	49	1
Cotton	K11	38	1
Paddy	Adt39	39	1

IV. Classification: Segmented databases

To improve predictive accuracy, databases can be segmented into more homogeneous groups. Then the data of each group can be explored, analyzed and modeled. Depending on the business question, segmentation can be done using variables associated with risk factors, profits or crop behaviors. Segments based on these types of variables often provide sharp contrasts, which can be interpreted more easily. Classification maps data into predefined groups or segments. Classification algorithms require that the classes be defined based on data attributes values. They often describe these classes by looking at the characteristics of data already known to belong to the classes. As a result, Government can more accurately predict the likelihood of a scheme based on the farmer's facilities in his land, which crop is suitable for land, how production can be increased, how crops are destroyed due to weather conditions.

DEFINITION

Given a database D = {t1,t2,…,tn } of tuples (items, records) and a set of classesC = {C1,…,Cm}, the classification problem is to define a mapping f : D _ C where each ti is assigned to one class. class Cj, contains precisely those tuples mapped to it that is, Cj ={ ti | f(ti) = Cj, 1_ i _ n, and ti _ D}

K Nearest Neighbors

When classification is to be made for new item using K Nearest Neighbors algorithm, its distance to each item in the training set must be determined. The new item is then placed in the class that contains the most items from the (K) closest set.

Input:
T //Training data
K //Number of neighbors
t //Input tuple to classify

Output:
c //class to which t is assigned.

Algorithm:
N = ø
//Find the set of neighbors, N,
for t For each d _ T do
If |N|_ K, then
N = N _ {d};
else
if u _ N such that sim(t,u) _ sim(t,d), then

begin
N = N – {u};
N = N _ {d};
end
//Find class for classification
C=class to which the most u _ N are classified;

For example, for crop paddy there can be three groups as first is for farmer with crop variety ponni. Similarly second one is for farmer with crop variety of ADT43, third ADT39.

TABLE 2: K Nearest Neighbors classification

CLASSIFICATION	Instances	% VALUE
Correctly Classified Instances	277	76.0989 %
Incorrectly Classified Instances	87	23.9011%
Kappa statistic		0.7278
Mean absolute error		0.0516
Root mean squared error		0.2156
Relative absolute error		29.2639 %
Root relative squared error		72.6678%

V. Correlation between Crops scheme

While studying scheme designing factor and scheme selection factor as a two variables simultaneously for a fixed population of farmers, government can learn much by displaying bivariate data in a graphical from that maintains the pairing. Such pair wise display of variables is called a scatter plot. When there is an increasing trend in the scatter plot, we say that the variables have a positive association. Conversely, when there is a decreasing trend in the scatter plot, we say that the variables have a negative association. If the trend takes shape along a straight line, then we say that there is a linear association between the two variables.

Going an example size of n and bivariate informational collection on these people or protests, the quality and straight connection between the two factors X and Y is estimated by the example relationship coefficient r, called the direct connection coefficient, measures the quality and the heading of a direct connection between two factors The straight relationship coefficient is once in a while alluded to as the Pearson item minute relationship coefficient out of appreciation for its designer Karl Pearson. The numerical equation for figuring r is:

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[\,n\Sigma x^2 - (\Sigma x)^2\,]\,[\,n\Sigma y^2 - (\Sigma y)^2\,]}}$$

The value of r is to such an extent that - 1 < r < +1. The + and – signs are utilized for positive direct connections and negative straight relationships, individually.

Positive correlation: If x and y have a solid positive straight relationship, r is near +1. A r estimation of precisely +1 demonstrates a flawless positive fit. Positive esteems show a connection amongst x and y factors to such an extent that as qualities for x increment, values for y likewise increment.

Negative correlation: If x and y have a solid negative straight relationship, r is near - 1. A r estimation of precisely - 1 shows a flawless negative fit. Negative esteems show a connection amongst x and y with the end goal that as qualities for x increment, values for y diminish.

No correlation: If there is no straight connection or a powerless direct connection, r is near 0. An incentive close to zero implies that there is an irregular, nonlinear connection between the two factors.

TABLE 3: Calculating standard deviation of total yield of a crop

Evaluation on test split	Calculated value
Correlation coefficient	0.6341
Mean absolute error	2218.1317
Relative absolute error	71.1355 %
Root relative squared error	108.9579 %
Total Number of Instances	124

TABLE 5.2 Standard deviation of total yield of a crop

Statistic	Value
Minimum	8
Maximum	96900
Mean	6784.629
StdDev	6269.8

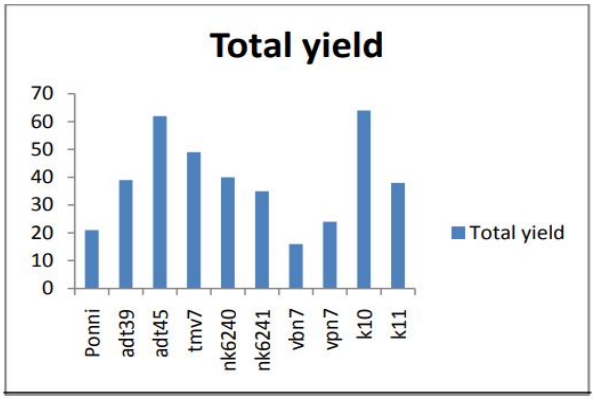


Fig 5.1 Visualizations of crop yielded more

CONCLUSION

- The study shows that adt39 yields more than other crop variety in paddy and k10 in cotton yielded more in one hectare.
- Cotton is a cash crop which is grown for sale to return a profit.
- The greater weight of maize is produced each year than any other grain to return a profit. In India, Maize is emerging as third most crop after rice and wheat.
- There is no linear correlation between total yields of the crops.

In the agriculture sector, data mining can help government to increase yield advantage mainly to support decision making, reliable and timely information on crop area, crop production and land use is of great importance to planners and policy makers for efficient agricultural development and for taking decisions on procurement, storage, public distribution, export, import and many other related issues to compete in the vend of crop pattern.

REFERENCES

1. J. Han and M. Kamber. Data Mining: Concepts and Techniques. Morgan Kaufmann, 3rd ed., 2011
2. Mr.A. B. Devale and Dr. R. V. Kulkarni “APPLICATIONS OF DATA MINING TECHNIQUES IN LIFE INSURANCE” IJDKP) Vol.2, ISSUE-4, July 2012.
3. Mr. Prof. Fakhreddine Karray “A Comparative Study of Data Clustering Techniques”
4. The estimates of the productivity and production of various crops were made based on the crop cutting experiment carried out with the joint efforts of the Departments of Economics and Statistics, Agriculture and Horticulture and Plantation crops, Retrieved from http://agritech.tnau.ac.in/pdf/2012/Season%20&%20Crop%20Report%202012.pdf.
5. Mr.A. B. Devale and Dr. R. V. Kulkarni “APPLICATIONS OF DATA MINING TECHNIQUES IN LIFE INSURANCE” IJDKP) Vol.2, ISSUE-4, July 2012.
6. Jain Rajni, Minz, S., V. Rama Subramaniam. 2009. "Machine learning for forewarning crop diseases". J. Ind. Soc. Agri. Stat. 63(1): pp. 97-107.
7. S. Veenadhari and B. Mishra, “Soybean Productivity Modelling using Decision Tree Algorithms”, 2011.
8. Jianlin Ji Dan, Qiu Chen, Jianping Chen, Li He Peng , 2010. "An improved decision tree algorithm and its application in maize seed breeding". Sixth Internation Conference on Natural Computation, held at Yantai, Shandon 10-12th January. pp. 117-121.
9. Crop Production in Ethiopia: Regional Patterns and Trends.
10. U. M. Fayyad, G. P. Shapiro, P. Smyth, and R. Uthusrsay, “Advances in Knowledge Discovery and Data Mining”, 3rd ed., 1996.
11. Olivia Parr Rud : “Data Mining, Modeling data for marketing risk, and Customer Relationship Management”, Wiley Publications 2003.
12. Kiran Mai, C., Murali Krishna, I.V., A.Venugopal Reddy, 2006. "Data Mining of Geospatial Database for Agriculture Related Application". Proc. of Map India. New Delhi.
13. Basak J., Sudharshan, A., Trivedi D., M.S.Santhanam. 2004. "Weather Data Mining Using Independent Component Analysis". J. of Machine Learning Research 5: pp. 239-253.
14. Patcharanuntawat, P., K. Bhaktikul, C. Navanugraha and T. Kongjun, 2007. Optimization for cash crop planning using genetic algorithm: A case study of upper Mun Basin, Nakhon Ratchasima province. 4th INWEPF Steering Meeting and Symposium, Paper 2-07: 2-13

15. Rajesh, D., 2011. Application of Spatial Data Mining for Agriculture. International Journal of Computer Applications, pp. 7-9.
16. Guidelines for the Validation And Verification of Quantitative and Qualitative Test Methods, National Association of Testing Authorities, 2012.
17. J. W. Mellor and P. Dorosh, “Agriculture and the Economic Transformation of Ethiopia”, 2010.
18. Ahsan Abdullah, Stephen Brobst, and Ijaz Pervaiz, “Learning Dynamics of Pesticide Abuse through Data Mining”, 2004.