Using Data Mining Techniques to Identify Crop Pattern

Dr.Yethiraj N.G, Punya H.N
Assistant Professor, Department of Computer Science, Maharani's Science College for Women, Bangalore

Abstract—Knowledge discovery in financial organization have been built to evaluate their operation and mainly to support decision making using knowledge as key factor. In this paper, we investigate the use of various data mining techniques for knowledge discovery in agriculture sector. We introduce different exhibits for discovering knowledge in the form of association rules, clustering, classification and correlation suitable for data characteristics.

Keywords — Agriculture crops, Association rules, Clustering, Classification and Data Mining.

INTRODUCTION

Data mining is emerging as an important research field. In this paper, we will discuss about the applications and techniques of Data mining in agriculture. 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. There are various data mining techniques such as K-Means, K-Nearest Neighbor (KNN), Artificial Neural Networks (ANN) and Support Vector Machines (SVM) which are used for very recent applications of Data Mining techniques. 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, identifyinginteraction terms and detecting nonlinear relationships. 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. Models that predict relationships and behaviors more accurately lead to greater profits and reduced costs.

Acquiring the Data

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 whichuninsured 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. 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 ismixture 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 itthat 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 isreached.

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, Kunigal. The collected data has been entered and analyzed using weka(machine learning

No.	cropname Nominal	surveynuml Numeric	cropvarie Non	totalyield Numeric
1	paddy	351.0	ponni	9290.0
2	paddy	359.0	ponni	10953.0
3	paddy	351.0	ponni	3144.0
4	paddy	351.0	ponni	16200.0
5	paddy	190.0	adt39	7962.0
6	paddy	206.0	ponni	7866.0
7	paddy	205.0	ponni	8264.0
8	paddy	261.0	ponni	8452.0
9	paddy	264.0	ponni	7941.0
10	paddy	351.0	ponni	3246.0
11	paddy	363.0	ponni	9877.0
12	paddy	350.0	ponni	6460.0
13	paddy	351.0	ponni	15930.0
14	paddy	204.0	adt39	8042.0
15	paddy	204.0	adt39	15850.0
16	paddy	207.0	ponni	15300.0
17	paddy	354.0	adt45	4122.0
18	paddy	365.0	adt45	3246.0
19	paddy	349.0	ponni	6570.0
20	paddy	352.0	adt45	4779.0
21	paddy	349.0	ponni	9650.0
22	paddy	326.0	ponni	8098.0
23	paddy	354.0	ponni	9558.0
24	paddy	174.0	adt39	7775.0

Fig 2.1 Visulaziation of Data

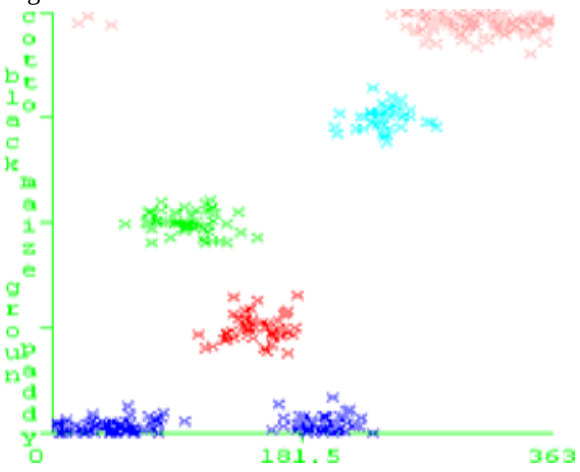


Fig 2.2 Culster Visulization using K Means

Retaining the Data

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 A and B are itemsets, A⊆I ,B⊆I ,A∩B=∅ .In support-confidence framework, each association rule has support and confidence to confirm the validity of the rule.
The support denotes the occurrence rate of an itemset in D, and the confidence denotes the proportion of data items containing B in all items containing A in D.

Sup(i) =
Count(i) /
Count(DBT)
Sup(A ⇒ B)
=Sup(AUB)
Conf (A ⇒ B) = Sup(AU B) / Sup(A)
When the support and confidence are greater than or equal to the pre-defined threshold 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
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 thenCk = Ck
U {I U J};

TABLE 3.1 Best rules found in apriori

Crop name	Crop varietyname	Support	Confidence
Maize	D-765	66	1
Paddy	Adt45	68	1
Ragi	Tmv7	49	1
Maize	Deccan- 107	40	1
Paddy	Adt39	36	1

Classification: Segmented Database

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 neighbours
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
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 Jenu gudu. Similarly second one is for farmer with crop variety of IR64, third IR20.

TABLE 4.1 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%

Correlation between Crop Schemes

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 a sample size of n and bivariate data set on these individuals or objects, the strength and linear

relationship between the two variables X and Y is measured by the sample correlation coefficient r, called the linear correlation coefficient, measures the strength and the direction of a linear relationship between two variables.

The linear correlation coefficient is sometimes referred to as the Pearson product moment correlation coefficient in honorof its developer Karl Pearson.

The mathematical formula for computing r is:

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

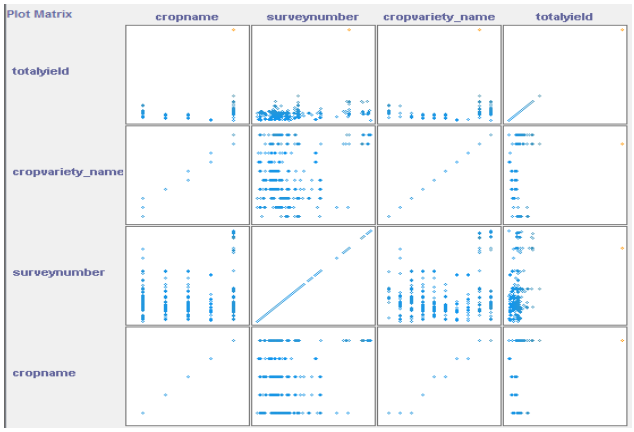
The value of r is such that -1 < r < +1. The + and – signs are used for positive linear correlations and negative linear correlations, respectively. Positive correlation: If x and y have a strong positive linear correlation, r is close to +1. An r value of exactly+1 indicates a perfect positive fit. Positive values indicate a relationship between x and y variables such that as values for x increase, values for y also increase. Negative correlation: If x and y have a strong negative linear correlation, r is close to -1. An r value of exactly -1 indicates a perfect negative fit. Negative values indicate a relationship between x and y such that as values for x increase, values for y decrease. No correlation: If there is no linear correlation or a weak linear correlation, r is close to 0. A value near zero means that there is a random, nonlinear relationship between the two variables.

TABLE 5.1 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 CROP

Statistic	Value
Minimum	8
Maximum	96900
Mean	6784.629
StdDev	6269.8



VI CONCLUSION

- The greater weight of maize is produced each year than any other grain to return a profit. In India, Maize isemerging 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.

REFERENCE

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 “A REVIEW OF DATA MINING TECHNIQUES IN INSURANCE SECTOR” Golden Research Thoughts Vol -1, ISSUE - 7 [January 2012]

3.Hongjun LU, Ling Feng and Jiawei Han, “Beyond Intratransaction Association Analysis: Mining Multidimensional Intertransaction Association Rules”, ACM Transactions on Information Systems, Vol. 18, October 2000.

4.Vaishali, A., Harsh, K., Anil, K.A, 2016, Performance Analysis of the Competitive learning Algorithms on Gaussian Data in Automatic Cluster Selection”, 2016 Second International Conference on Computational Intelligence &Communication Technology.

5.Srinivas, K., Kavita, R.B., Govardhan, A., 2010 "Applications of Data Mining echniques in Healthcare and Prediction of Heart attacks" International Journal on Computer Science andEngineering", 2(2), pp.250-255
6.Salim, D., Suzan Mishol., Daniel, S.K., Dina M., Anael S., 2013, "Overview Applications of Data Mining in HealthCare: The Case Study of Arusha Region" International Journal of Computational Engineering Research, 3(8), pp. 73-77.
7.Darcy, A. D, Nitesh V.C., Nicholas B, 2008, "Predicting Individual Disease Risk Based on Medical History" CIKM '08 Proceedings of the 17th ACM conference on Information and Knowledge management, pp. 769-778.
8. Eibe F, Mark A.H, IanH.W. 2016, "The WEKA Work bench. Online Appendix for Data Mining: Practical Machine Learning Tools and Techniques", Morgan Kaufmann, Fourth Edition.
9. Lichman, M., 2013, "UCI Machine Learning Repository" [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.
10.Patil, B. U., Ashoka, D. V., & Prakash, B. V. A. (2023). Data Integration Based Human Activity Recognition using Deep LearningModels. Karbala