International Journal of New Computer Architectures and their Applications (IJNCAA) 5(2): 48-52
The Society of Digital Information and Wireless Communications, 2015 (ISSN: 2220-9085)

Application of Machine Learning Techniques for Yield Prediction on Delineated Zones in Precision Agriculture

Anshal Savla[1], Himtanaya Bhadada[1], Parul Dhawan[1], Vatsa Joshi[1],

[1]Department of Information Technology, NMIMS' MPSTME,
Mumbai, India

anshalsavla@gmail.com, himtanaya@gmail.com, paruldpsr@yahoo.com, vatsajoshi94@gmail.com

Abstract. Precision agriculture is the implementation of the recent technology in agriculture. Huge amount of data is collected in agriculture and various techniques of data mining are used to make efficient use of it. In this paper, we have discussed how with the help of both, clustering and classification algorithms, the crop suitable for a particular piece of land can be determined. Management zone delineation is a key task in this. From a data-mining point of view this comes down to variant of spatial clustering which has a constraint of keeping the resulting clusters spatially mostly contiguous. We analyze the need to discretize and normalize the data set and the various techniques that are used for the same. Further, a comparative analysis of the algorithm is shown where it can be seen which algorithm is best suited. We also talk about the future scope of the same and how these could actually be implemented in the real life scenarios.

KEYWORDS

Discretization, Normalization, Clustering, Classification, Precision agriculture, Zone delineation.

1 INTRODUCTION

In recent years precision agriculture [1] has gained a lot of attention due to enormous possibilities it can open up in the field of agriculture. In one of our previous paper titled "Survey of Zone Tessellation Techniques for Defined Parameters in Precision agriculture", we had applied various clustering algorithms on a dataset and used zone delineation for predicting the yield of crops in a particular area. DBSCAN and ICEAGE algorithms had the best time complexity. In another paper titled "Survey of classification algorithms for formulating yield prediction accuracy in precision agriculture" we have obtained the graphs and results presented in the paper. These are the result of the application of various classification algorithms on the dataset for predicting the yield of the crop. Further it was seen that Bagging algorithm gives the least error in predicting the seeds for the crop for a particular year.

The data set that we have used comprises of the following attributes: Area harvested, Seed, Yield and Production. The data is collected for the soybean crop. Yield is the actual generation of seed from the soybean crop, area harvested is the amount of the crop collected in a season and production is the quantity produced and actually harvested for the soybean crop. The data range over 53 years, starting from 1961 to 2013.

In the paper, we have provided a combined approach of both the techniques of zone delineation and prediction. First we divide the entire farm in zones using clustering technique and then we apply classifying technique for yield prediction of each zone. In section 2 of the paper we discuss zone delineation with respect to clustering techniques, in section 3 data preprocessing methods (normalization, discretization) are explained.

2 LITERATURE SURVEY

2.1 Normalization

The technique in which data is organized in a database is Normalization [2]. The process involves number of steps. It refactors the tables a table into number of smaller tables that are less redundant. No information is lost during the process and foreign keys are defined in old tables that refer to the primary keys in the new ones. Normalization is done in order to isolate the data so that the changes done in the attributes of one table can be sent to the entire database easily. Creation of tables and defining the relationships between them is done in order to protect the data.

Data normalization is a useful concept in organizing the data set. Without it the data system can become slow, inefficient and inaccurate. Normalization aims to organize the data into logical groups where every group effectively describes the small part of data. Also it becomes easy to modify the data in large database. This is because the change that is done is only at one place. Ease of access and quick manipulation of data is done by normalizing the date set.

$$Z_i = \frac{x_i \text{ - } \min(x)}{\max(x) \text{ - } \min(x)} \text{ ....(1)}$$

where,

x_i= current value
z_i=normalized value
x=column vector

2.2 Discretization

Before data-mining, often data preprocessing is required. (ex: normalization, discretization)
When using data mining we usually work on large sets of data wherein an attribute value can vary over an extremely large range. It becomes necessary to reduce the number of values of a continuous attribute and divide the range into discrete intervals.

Equal distance discretizer:
The equal distance discretizer [3] is a simple, static and unsupervised method that divides the entire range of an attribute that has to be discretized into N equal intervals where the user specifies the value of N. The algorithm computes the minimum (Vmin) and maximum (Vmax) values for an attribute and divides it into k parts: intervals=(Vmax+Vmin)/K where 'K' is provided by the user and boundaries = Vmin+(
i * interval) for the i = 1...k-1 boundaries.

The limitations of this method are:

1. It is a parameterized method that requires input from the user.
2. The data values are not equally distributed over the intervals.

Equal frequency discretizer:
Equal frequency discretizer is a static, unsupervised and parametric method. It finds the minimum and maximum value attributes and then arranges all the values in increasing order. It then divides the entire range (n values) of an attribute in such a way that each interval has almost the same number of samples (interval= n/k) [4]. Thus, it overcomes the shortcomings of Equal distance discretizer.

Chi2
The chi2 algorithm is comprised of two parts. In the first part, the algorithm starts with computing a high significance level (sigLevel) for each numeric attribute that has to be discretized and then all the attributes are sorted according to its sigLevel [5].
It then calculates the Chi^2 for every pair of adjoining intervals. In the second part of the algorithm adjacent intervals with the lowest Chi^2[6] values are merged. This process continues until Chi^2 values of all interval pairs exceed the parameter determined by the sigLevel. The entire process is repeated for decreasing values of sigLevel until an inconsistency rate is exceeded in the discretized data.

$$\mathbf{x^2} = \sum_{i=1}^{2}\sum_{j=1}^{k}\frac{(A_{ij}-E_{ij})^2}{E_{ij}} \text{ ....(2)}$$

where[15],
K = number of classes,
A_{ij} = number patterns in the ith interval, jth class,
R_i = number patterns in the ith interval
C_j = number of patterns in the jth class
N = total number patterns
E_{ij} = expected frequency of A_{ij} = R_i * C_j / N

Class attribute contingency coefficient discretization

This discretization method is proposed by Lee et al [7] which is based on the concept of class attribute contingency coefficient. This supervised and top down method is very efficient as different attributes are considered. Discretization of continuous data is done by measuring the value of CACC. The minimum and the maximum value of each attribute is determined and then all the values of each attribute is sorted in the ascending order. The attributes are partitioned according to the maximum CACC value into intervals.

2.3 Zone Delineation with respect to Clustering Techniques

In clustering techniques, zone delineation is a vital aspect. It helps us understand which part of the land is best suited for a particular type of crop. As opposed to homogenous crop selection methods like in traditional agriculture, we can have a heterogeneous crop selection using zone delineation in precision agriculture.

In DBSCAN and ICEAGE algorithms, the agricultural data attribute (k) is responsible for the density of the redundant cluster.

In our data set, there are fifty-three data entities and if we label each entity as ‘k’, then the density of the cluster increases since every individual cluster will be one single value. However if we have only three clusters of approximately seventeen data values each, then the density of the cluster will be very low.

Ideally the value of ‘k’ should be between 3 to 10.

2.4 Application of Classification Over Individual Zone

In previous papers, various classification algorithms are discussed. After analyzing all the algorithms from previous papers we came to a conclusion that bagging is the best algorithm for predicting the yield for data set used in previous papers. Therefore, bagging algorithm is used to predict the yield of crop over individual zones.

3 PROPOSED MODEL

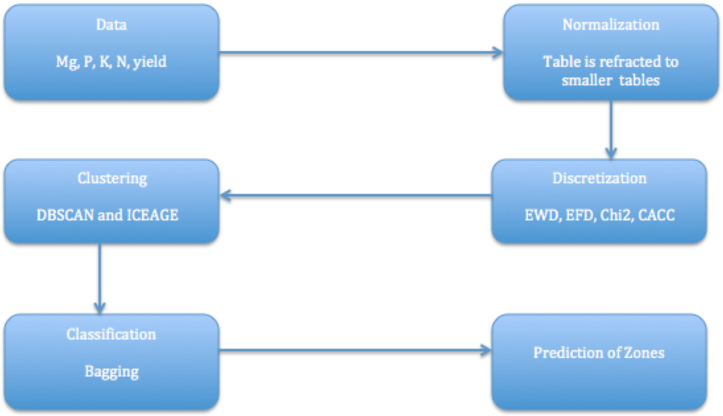

Figure 1: Block diagram

Step 1: Import data set

Step 2: The variables used in the data set are transformed into a specific range. This brings the data set into a consistent state and so anomalies are avoided. Once the data set is normalized redundancy is reduced and also managing of data becomes easier.

Further after refactoring is done old table are assigned the foreign keys that refer to the primary keys of the new ones.

Step 3: Continuous data set is then discretized. We have used Equal Distance Discretizer, Equal Frequency Discretizer, Class Attribute Contingency Coefficient and Chi2 discretization techniques.

Table 2: Comparison of discretization techniques

Characteristics	**Equal Distance Discretizer**	**Equal Frequency Discretizer**	**Class Attribute Contingency Coefficient**	**Chi2**
Supervised/ Unsupervised	Unsupervised	Unsupervised	Supervised	Supervised
Top Down/ Bottom Up	Top Down	Top Down	Top Down	Bottom Up
Parametric	Yes	Yes	No	Yes
Incremental	No	No	Yes	Yes
Static/ Dynamic	Static	Static	Static	Static

Step 4: Clustering is then applied to the existing data set. The clustering technique used is DBSCAN, as the analysis done in previous papers showed that DBSCAN has the best complexity.

Step 5: Classification method is applied on the clustered data set. The classification technique used is bagging. Bagging is used since it was best suited for yield prediction as shown in previous papers.

4 EXISTING COMPARATIVE ANALYSIS

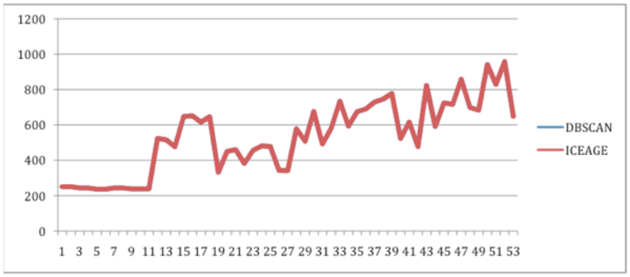

Figure 3: Comparison of DBSCAN and ICEAGE complexity

From previous papers we conclude that DBSCAN [8] and ICEAGE [9] have the best time complexity.

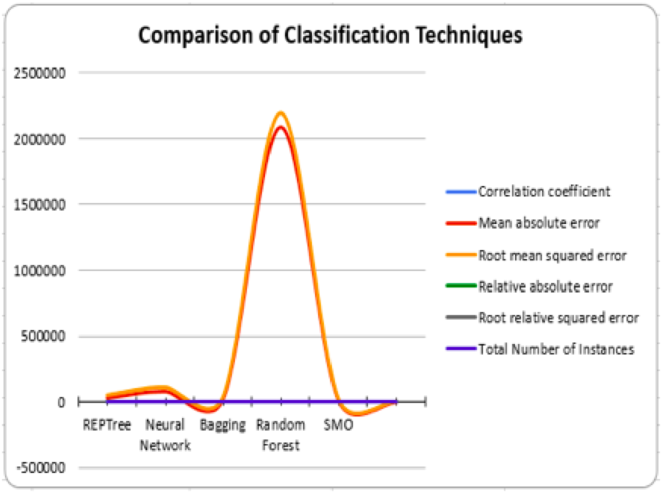

Figure 4 : Graphical view of accuracy parameters

From previous papers we conclude that bagging [10] is the most efficient algorithm among REPTree [11], Neural Network [12], Random Forest [13] and SVM [14] since it has the least error deviation.

5 FUTURE WORK

In the future we plan to work on doing comparative analysis of various different discretization techniques and choose the best fit for the proposed model.

6 CONCLUSION

In this paper we studied and described a framework that can help us analyze and understand the yield of the crop for a designated zone based on the density of attributes. In future work we would fully develop a DSS that could provide decisions on the type of crop for each zone based on NPK parameters.

7 REFERENCES

1. Bongiovanni, Rodolfo, and Jess Lowenberg-DeBoer.: Precision agriculture and sustainability. Precision Agriculture 5.4 (2004): 359-387.
2. Stalikas, Constantine D., George A. Pilidis, and Stella M. Tzouwara-Karayanni.: Use of a sequential extraction scheme with data normalisation to assess the metal distribution in agricultural soils irrigated by lake water. Science of the total environment 236.1 (1999): 7-18.

3. Lim, Meng-Hui, and Andrew Beng Jin Teoh.: An analytic performance estimation framework for multibit biometric discretization based on equal-probable quantization and linearly separable subcode encoding. Information Forensics and Security, IEEE Transactions on 7.4 (2012): 1242-1254.
4. R. Chaves, J. Ramirez and J.M. Gorriz,: Integrating discretization and association rule-based classification for Alzheimer's disease diagnosis. Expert Systems with Applications, pp. 1571–1578, 2013
5. Garcia, Salvador.: A survey of discretization techniques: Taxonomy and empirical analysis in supervised learning." Knowledge and Data Engineering, IEEE Transactions on 25.4 (2013): 734-750.
6. Su, Chao-Ton, and Jyh-Hwa Hsu.: An extended chi2 algorithm for discretization of real value attributes. Knowledge and Data Engineering, IEEE Transactions on 17.3 (2005): 437-441.
7. MacQueen, James.: Some methods for classification and analysis of multivariate observations. Proceedings of the fifth Berkeley symposium on mathematical statistics and probability. Vol. 1. No. 14. 1967.
8. Zhou, Aoying,: Approaches for scaling DBSCAN algorithm to large spatial databases. Journal of computer science and technology 15.6 (2000): 509-526.
9. Guo, Diansheng, Donna J. Peuquet, and Mark Gahegan.: ICEAGE: Interactive clustering and exploration of large and high-dimensional geodata. GeoInformatica 7.3 (2003): 229-253.
10. Bauer, Eric, and Ron Kohavi.: An empirical comparison of voting classification algorithms: Bagging, boosting, and variants. Machine learning 36.1-2 (1999): 105-139.
11. Mohamed, W. Nor Haizan W., Mohd Najib Mohd Salleh, and Abdul Halim Omar.: A comparative study of reduced error pruning method in decision tree algorithms. Control System, Computing and Engineering (ICCSCE), 2012 IEEE International Conference on. IEEE, 2012.
12. Gorman, R. Paul, and Terrence J. Sejnowski.: Analysis of hidden units in a layered network trained to classify sonar targets. Neural networks 1.1 (1988): 75-89.
13. Gislason, Pall Oskar, Jon Atli Benediktsson, and Johannes R. Sveinsson.: Random forests for land cover classification. Pattern Recognition Letters 27.4 (2006): 294-300.
14. Vishwanathan, S. V. M., and M. Narasimha Murty.: SSVM: a simple SVM algorithm. Neural Networks, 2002. IJCNN'02. Proceedings of the 2002 International Joint Conference on. Vol. 3. IEEE, 2002.
15. Liu, Huan, and Rudy Setiono.: Chi2: Feature selection and discretization of numeric attributes. 2012 IEEE 24th International Conference on Tools with Artificial Intelligence. IEEE Computer Society, 1995.