iJCEM
International Journal of Core Engineering & Management

ISSN: 2348 9510

International Journal Of Core Engineering & Management (IJCEM)

Volume 1, Issue 5, August 2014

Indian Agriculture Land through Decision Tree in Data Mining

Kamlesh Kumar Joshi, M.Tech(Pursuing 4'th Sem)

Laxmi Narain College of Technology, Indore (M.P) India

k3g.kamlesh@gmail.com

9926523514

Pawan Patidar, Assistant Professor

Laxmi Narain College of Technology, Indore (M.P) India

pawanpatidar4u@yahoo.co.in

9926433364

ISSN: 2348 9510

International Journal Of Core Engineering & Management (IJCEM)

Volume 1, Issue 5, August 2014

Abstract

The decision tree is one of the common modeling methods to classify. Firstly, this paper introduces the concept of Classification and the method of the decision tree. Then, this paper analyses the data of rural labor, arable land area and the gross output value of agriculture about 10 cities of India based on the decision tree, and adopts clustering analysis method to discretize continuous data during the process of data miming in order to subjectivity comparing to the traditional classification methods.

Finally, generating the decision tree of our agriculture, thereby gaining the spatial classification rules and analyzing the rules.

Keywords: classification rule, Decision Tree, discretization , clustering analysis ,generalizing idea, agriculture.

I. Introduction

Agricultural land grading is the integrated assessment of the agricultural land in the administrative region, which is reflected by some natural and socioeconomic factors. Traditional methods for gradating agricultural land are mainly factor method, comparable plot method and modification method [1]. On the other hand, the materials about land information would probably be incomplete. And the traditional methods cannot perform well in dealing with the misdata and missing data. Furthermore, the traditional methods mainly depend on experiential knowledge, so that they don't have the ability of self-learning and can't dispose of the qualitatively described variables well.

Decision tree is one of the classification methods, and it is used widely in data mining. And it has been broadly applied in information extraction from remote sensing image, disaster weather forecasting, correlation analysis of environmental variables, and so on [2,3,4]. The decision tree analysis method has its own advantage in solving the above problems that traditional methods cannot solve. Moreover, agricultural land grading can be seen as the classification of the mixed spatial data which is derived from the quantization of the factors that are impacting the land quality, and the result is the agricultural land grade. Therefore, in order to overcome the limitation of traditional methods, our study applied the decision tree analysis method into agricultural land grading and constructed the decision tree model in M-language based on MATLAB.

II. The Basic Concepts Of Classification

Rule Mining

The classification rule mining belongs to the scope of Data Mining. Each object in the presumptive data base (each tuple in RDB viewed as one object) belongs to a given class which is confirmed by the attribute of identifiers, and classification is the process of allotting data of the database to the given class. There exists a large number of arithmetic in classification, for instance, Fayyad U M, Piatetsky-SShapiro, and Smyth P, Eds, 1996; Quinlan J R,1986; Quinlan J R, 1990; Safavian S R and langrebe D, 1991. Common statistic method can only effectively deal with continuous data or discrete ones (Quinlan J R, 1990), but decision tree can deal with both numerical data and symbolic data. Many statistic classification methods and neural net methods use equations to denote information, while decision tree transfers information into rules, it is crucial for decision, because each person would like to make decisions according to comprehensible information, and be unwilling to do it according to "black-box". Spatial classification rule is different from many other classification methods, the formal one only considers relational data, while the latter one also needs to consider spatial data, for instance, geographical data contains the description of both spatial object and non-spatial object. The description of non-spatial object can be restored in traditional relational data base, and needs to set a attributive pointer pointing at the spatial description of the object. In the process of spatial classification, searching the rules for dividing object sets to different classifications not only needs to use the relationship between the attributes of the classified objects, but also needs the relationship between the classified objects and other objects in the data base.

III. The Method Of Decision Tree

Decision tree looks like a real tree, it adopts the superincumbent strategy to distribute given object into small data sets, and in these small data sets, the leaf crunodes usually connect to only one category.

The basic concepts of decision tree:

1. The decision tree is constructed by the superincumbent and divide-and-conquer mode

2. All attributes are categorical, and the attributes of continuous value must to be discretized in advance

3. At the beginning, all disciplinal samples are on the root

4. The samples on the nodes recursively based on the decided partition of the attributes

5. The selection of attributes is based on the heuristic or statistical measurement

The condition of stop division:

1. All samples from the given nodes belong to the same category

2. No character to be divided further --- to classify the leaf crunodes by majority vote

3. There's no sample on the given nodes

Algorithm: Generate_decision_tree

Input:

●data division D: disciplinal metagroup and the collection of category marks they refer to

●attibute_list: the collection of candidate charactors

●Attribute_selection_method: an assured "best" process of schismatical discriminant which can plot metadata group to classes. This criterion consists of schismatical attributes and schismatical points or schismatical subsets.

Output: a decision tree

Method:

```
1. Create node N
2. If samples are all in the same class C then
3. return N as a leaf node, marked as class C
4. If attribute_list is vacant then
5. Return N as leaf node, marked as the majority class of the samples; // majority vote
6. Use attribute_selection_method (D, attribute_list) find out the "best" splitting_criterion
7. Use splitting_criterion to mark N
8. If splitting_attribute is discrete and allows multiprogramming then //un restrict to double-branch tree
9. Attribute_list ← attribute_list – splitting_attribute; // delete plot attributes
10. for splitting_criterion, each result j //plot meta group and produce subtree for each partition
11. Suppose j D is the set of metadata in D be up to the result; // one classification
12. If j D is vacant then
13. Plus one leaf, marked as the majority of j D
14. Else plus a node which returned from
Generate_decision_tree ( j D , attribute_list) to N;
15. Return N
```

The method of decision tree requires all characters are classified. So the character of continuous value should be pre- discretized. On selecting the discrete method, people usually classify them on the conceptual level according to experience, which is subjective and the researcher is required to have plenty of background knowledge on the study data. In practice, the problems on the levels of division, how to definite splitting point, and the division of regions are usually solved by the experience and long-term experimental opinion finding out optimal value to confirm, but the model which can handle these kinds of problems on the level of knowledge rarely exists. In this paper, we make cluster analysis by SPSS software firstly, and then generalize the classified conceptions in order to achieve to divide the data into different classes and levels.

Iv. Data Mining And Analysis Of Our Agriculture

The experimental data are from the total output value of an annual agriculture production, there lists data of 10 provinces and cities' (such as, Delhi, Chandigarh, Nasik, Kanpur, Surat, and so on) rural labor, acreage of plantation and total output value of agriculture. The detail data are listed in Table 1:

TABLE 1. TOTAL OUTPUT VALUE OF AN ANNUAL AGRICULTURE PRODUCTION

City	Rural labor	Arable land Area	Gross Agriculture production
Delhi	67.7	399.5	176.58
Srinagar	79.4	426.1	156.17
Chandigarh	1635.5	6517.3	1505.94
Surat	639.9	3645.1	359.15
Chennai	512.4	5491.4	534.39
Kanpur	633.0	3389.7	969.79

Jaipur	517.0	3953.2	666.47
Hyderabad	760.3	8995.3	736.34
Bhopal	76.3	290.0	206.78
Nasik	1531.5	4448.3	1849.18

Use SPSS to classify the data of rural labor, acreage of plantation and total output value of agriculture (cluster analysis use association method, the calculation of distance use square Euclidean Distance), and the results are listed in Table 2:

TABLE 2. THE RESULT OF CLUSTER ANALYSIS

City	Rural labor Class	Arable land Area Class	Gross Agriculture Production Class
Delhi	1	1	1
Srinagar	1	1	1
Chandigarh	2	3	2
Surat	1	2	1
Chennai	1	3	1
Kanpur	1	2	2

Jaipur	1	2	1
Hyderabad	1	3	1
Bhopal	1	1	1
Nasik	2	2	3

To carry out generalization conceptual process on the results of the classification, that is:

Rural labor:

1 → few;

2 → medium;

3 → much

Arable land: 1 → small;

2 → medium;

3 → large

Gross agriculture production:

1 → low;

2 → medium;

3 → high

The total agricultural output information that has been generalized is shown in Table 3:

TABLE 3: TOTAL AGRICULTURAL OUTPUT INFORMATION

AFTER GENERALIZATION

City	Rural labor Class	Arable land Area Class	Gross Agriculture Production Class
Delhi	Few	Small	Low
Srinagar	Few	Small	Low
Chandigarh	Medium	Large	Medium
Surat	Few	Medium	Low
Chennai	Few	Large	Low
Kanpur	Few	Medium	Medium
Jaipur	Few	Medium	Low
Hyderabad	Few	Large	Low
Bhopal	Few	Small	Low
Nasik	Medium	Medium	High

From the Table 2, the cluster analysis of the labor situation and the total Agricultural output, we know that total Agricultural output has much concern with the number of rural labor, and if the number is high, the total Agricultural output is high, and vice versa.

From the analysis above, what we got from the rule corresponds to the current agricultural situation in India. During discretization process of continuous data, we find that cluster analysis method well avoids the subjective effects arise from categorization by experience, and reflects the reality.

V. Conclusion

As a new analysis method and approach in finding the potential information in mass data, Data Mining has attracted much attention all over the world. Among them, Decision Tree with high data-processing efficiency and easily-understood characteristics becomes much more popular and has already been widely used in many fields, for example, speech recognition, medical treatment, model recognition and expert system, etc. And it includes many methods, and each method has its character, so we should chose the best method according to the specific data category. In addition, methods are complementary with another to combine into a whole system and all of them aim to process and refine the potential information. In conclusion, How to find the better method to pre-process the date is long and extensive pursuit by all of us.

VI. References

[1] Lv An-ming, Li Cheng-min, Lin Zong-jian, Wang Jia- ao．GIS Attribute Data Mining based on Statistic Inducfion[J]． Journal of Zhengzhou Institute of Surveying and Mapping, 2001，18(4)： 290-293

[2] Mehmed Kantardzic．Data Mining – Concepts, Models, Techniques and Algorithms[M]． Shan Si-qing etc ranslate． BeiJing: Tsinghua University Press, 2003

[3] Margaret H Dunham． Data Mining[M]． Guo Chong-hui, Tian Zhan-feng etc translate． BeiJing: Tsinghua University Press, 2003

[4] Quinlan J R． Induction of decision trees． Machine Learning． 1986： 1—356

[5] Cai Zhi-hua，Li Hong，Hu Jun． Decision Tree Algorithm to Spatial Classification Rule Mining[J]． Computer Engineering，2003， 29(11)：74-75，118

[6] Li Qiang．A Comparative Study on Algorithms of Constructing Decision Trees —— ID3, C4. 5 and C5.0[J]．Journal of Gansu

Sciences, 2006，18(4)：84-87

[7] Yang Xue-bin, Zhang Jun． Decision Tree and Its Key

Techniques[J]．Computer Technology and Development, 2007，17(1)：43-45