ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 3, March 2015

Application of Classification Technique in Data Mining for Agricultural Land

M.C.S. Geetha

Assistant Professor, Department of Computer Applications, Kumaraguru College of Technology, Coimbatore, India

Abstract: Data mining is the process of finding new patterns. Classification is the method of generalizing known structure to pertain to new data. Classification using a decision tree is achieved by routing from the root node until arriving at a leaf node. To model classification process, decision tree is used. Also there are many classification algorithms available in literature but decision trees is the most commonly used because of its ease of implementation and easier to understand compared to other classification algorithms. This paper begins with the basic concepts of Classification and the method of the decision tree. Then, this paper analyses the data of arable land area, rural labor and the gross output value of agriculture about 10 cities of India based on the decision tree, and implement clustering analysis method to discrete data during the process of data mining compared to the traditional classification methods. Finally, carry out generalization conceptual process from the results of the classification [1].

Keywords: classification concepts, Decision Tree, clustering, agriculture, production data set.

I. INTRODUCTION

Agricultural land grading is the integrated assessment of the agricultural land in the administrative region, which is reflected by some socioeconomic and natural factors. Traditional methods for radiating agricultural land are mainly factor method, comparable plot method and modification method [2]. On the other hand, the materials about land information would perhaps be incomplete. And the traditional methods cannot carry out 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 [3,4]. The decision tree analysis method has its own advantage in solving the above problems that traditional methods cannot solve. 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 score. Therefore, in order to overcome the constraint 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.

The paper is organized as follows: Chapter 2 discusses the basic concepts of classification. Chapter 3 discusses the method of decision tree. Chapter 4 discusses the data mining analyses in agriculture. Chapter 5 discusses the conclusion.

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. [5] Common statistic method can only effectively deal with continuous data or discrete ones [6] 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 separating object sets to different classifications not only needs to apply 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 trees are trees that classify instances by sorting them based on characteristic values. Each node in a decision tree represents a characteristic in an instance to be classified, and each branch signifies a value that the node

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 3, March 2015

can imagine. Instances are classified starting at the root node and sorted based on their characteristic values [7]. Decision tree rules provide model transparency so that a user can know the basis of the model's predictions, and therefore, be comfortable performing on them and explaining them to others[8].

Decision tree algorithm is a data mining induction techniques that recursively partitions a data set of records by means of depth-first greedy approach [9] or breadth-first approach [10] until all the data items belong to a meticulous class. A decision tree structure is made of root, leaf and inner nodes. The tree structure is used in classifying unfamiliar data records. At each inner node of the tree, a decision of best opening is made using impurity measures [11]. The tree leaves is made up of the class labels which the data items have been group.

Decision tree classification technique is performed in two phases: tree building and tree pruning.

Tree building is done in top-down manner. In this phase, the tree is recursively partitioned till all the data items belong to the same class label. It is very tasking and computationally intensive as the training data set is traversed repeatedly.

Tree pruning is done is a bottom-up manner. It is used to develop the prediction and classification accuracy of the algorithm by minimizing over-fitting (noise or much detail in the training data set) [12,13]. Tree pruning is less tasking compared to the tree growth phase as the training data set is scanned only once.

The basic concepts of decision tree:

1. The decision tree is constructed by the superincumbent and divide-and-conquer mode
2. All attributes are definite, and the attributes of permanent 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 set of category marks they refer to
- attibute_list: the collection of candidate characters
- Attribute_selection_method: an assured "best" process of schismatical discriminate which can plot metadata collection to classes. This principle 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-discriminated. On selecting the distinct method, people usually organize 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 partition, how to distinct splitting point, and the division of regions are usually solved by the experience and long-term experimental opinion finding out optimal value to prove, but the demonstration which can handle these kinds of problems on the level of knowledge rarely exists. In this paper, cluster analyses by SPSS (Statistical Package for the Social Sciences) software is done first and then simplify the classified conceptions in order to achieve to divide the data into different classes and levels.

IV. DATA MINING ANALYSES IN AGRICULTURE

The investigational data are from the total output value of a yearly agriculture production, there lists data of 10 region and cities' (such as Jodhpur, Hyderabad, Indore, Akola, and so on) rural labor, acreage of plantation and total production value of agriculture.

The data are listed in Table 1:

TABLE 1.
TOTAL VALUE OF A YEARLY AGRICULTURE PRODUCTION

City	Rural Labor	Arable land Area	Gross Agriculture production
Jodhpur	517.1	3953.1	666.48
Hyderabad	760.4	8995.4	736.35
Indore	76.4	290.1	206.79
Akola	1531.6	4448.4	1849.19
Delhi	67.8	399.6	176.59
Jammu	79.5	426.2	156.18
Chandigarh	1635.6	6517.4	1505.95
Ahmedabad	639.8	3645.2	359.16
Coimbatore	512.5	5491.5	534.40
Lucknow	633.1	3389.8	969.80

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 3, March 2015

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

TABLE 2.
THE RESULT OF CLUSTER ANALYSIS

City	Rural Labor Class	Arable land Area Class	Gross Agriculture Production Class
Jodhpur	1	2	1
Hyderabad	1	3	1
Indore	1	1	1
Akola	2	2	3
Delhi	1	1	1
Jammu	1	1	1
Chandigarh	2	3	2
Ahmedabad	1	2	1
Coimbatore	1	3	1
Lucknow	1	2	2

To carry out generalization conceptual process on the results of the categorization:

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

TABLE 3:
OVERALL AGRICULTURAL OUTPUT INFORMATION AFTER GENERALIZATION

City	Rural labor	Arable land Area	Gross Agriculture production
Jodhpur	Few	Medium	Low
Hyderabad	Few	Large	Low
Indore	Few	Small	Low
Akola	Medium	Medium	High
Delhi	Few	Small	Low
Jammu	Few	Small	Low
Chandigarh	Medium	Large	Medium
Ahmedabad	Few	Medium	Low
Coimbatore	Few	Large	Low
Lucknow	Few	Medium	Medium

From the Table 2, the cluster investigation of the labor situation and the overall agricultural output, it shows that the overall agricultural output has much concern with the number of rural labor, and if the number is high, the overall agricultural output is high, and vice versa.

From the analysis above, it indicates from the rule corresponds to the current agricultural position in India. During discretization process of permanent data, it proves that cluster analysis method avoids the personal effects arise from classification by practice, and reflects the reality.

A. 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 techniques, and each method has its character, so chose the best method according to the specific data type. In addition, the techniques are complementary with another to combine into a whole system and all of them aim to process and refine the potential information.

REFERENCES

[1] Li, L., Zhang, X. (2010). Study of data mining algorithm based on decision tree. International Conference On Computer Design And Applications Vol. 1.
[2] 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
[3] Mehmed Kantardzic. Data Mining – Concepts, Models, Techniques and Algorithms[M]. Shan Si-qing etc ranslate. BeiJing: Tsinghua University Press, 2003
[4] Margaret H Dunham. Data Mining[M]. Guo Chong-hui, Tian Zhan-feng etc translate. BeiJing: Tsinghua University Press, 2003
[5] Quinlan, J. R. (1986). Induction of decision trees. Machine Leaning, vol (1), pp.81-106
[6] Quinlan, J. R. (1987). Simplifying decision trees, International Journal of Machine Studies, number27, pp. 221-234.
[7] Decision Trees for Business Intelligence and Data Mining: Using SAS Enterprise Miner "Decision Trees— What Are They?"
[8] Lior Rokach and Oded Maimon,"Data Mining with Decision Trees: Theory and Applications(Series in Machine Perception and Artificial Intelligence)", ISBN: 981-2771-719, World Scientific Publishing Company, , 2008.
[9] Hunt, E.B., Marin. and Stone,P.J. (1966). Experiments in induction, Academic Press, New York
[10] Shafer, J., Agrawal, R., and Mehta, M. (1996). Sprint: A scalable parallel classifier for data mining. Proceedings of the 22nd international conference on very large data base. Mumbai (Bombay), India
[11] Quinlan, J. R. (1993). C45: Programs for Machine Learning. Morgan Kaufmann, San Mateo, CA.
[12] Mehta, M., Agrawal, R., and Rissanen, J. (1995). MDL-based decision tree pruning. International conference on knowledge discovery in databases and data mining (KDD-95) Montreal, Canada
[13] Mehta, M., Agrawal, R., and Rissanen, J. (1996). SLIQ: A fast scalable classifier for data mining. In EDBT 96, Avignon, France

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 3, March 2015

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

[15] 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

[16] Yang Xue-bin, Zhang Jun. Decision Tree and Its Key Techniques [J]. Computer Technology and Development, 2007, 17(1) : 43-45

BIOGRAPHY

M.C.S. Geetha is an Assistant Professor in the Department of Computer Applications, Kumaraguru College of Technology, Coimbatore. She received Master of Computer Applications (MCA) degree in 2004 from P.S.G.R.Krishnnammal College for Women, Coimbatore, India. She received M.Phil (Computer Science) in 2006 from Bharathiyar University. Her research interest is Data Mining. She has published papers in International Journal.