Research on Rough Set and Decision Tree Method Application in Evaluation of Soil Fertility Level

Guifen Chen[1], Li Ma[1]
[1]College of Information and Technology Science, Jilin Agricultural University,Chang Chun，Jilin，China
guifchen@163.com,Mary19801976@sohu.com

Abstract. Clustering, rough sets and decision tree theory were applied to the evaluation of soil fertility levels ,and provided new ideas and methods among the spatial data mining and knowledge discovery. In the experiment, the rough sets - decision tree evaluation model establish by 1400 study samples, the accuracy rate is 92% of the test. The results show :model has good generalization ability; using the clustering method can effectively extract the typical samples and reducing the training sample space; the use of rough sets attribute reduction, can remove redundant attributes, can reduce the size of decision tree decision-making model, reduce the decision-making rules and improving the decision-making accuracy, using the combination of rough set and decision tree decision-making method to infer the level of a large number of unknown samples.

Keywords: rough set; decision tree; clustering； data mining; soil evaluation; productivity grade

1 Introduction

In precision agriculture, soil fertility evaluation is the basis for land management, which not only can assess the level of land productivity, and can guide the rational development and utilization.

From the data mining point of view, fertility is essentially within the classification prediction. Characteristics of the decision tree method are those: decision tree is a better classification method can handle the non-linear data and describe the data with generating speed and set up an intuitive tree structure [1]. Rough set theory has some advantages in the processing of data and eliminating redundant information and dealing with uncertain information， so widely used in data preprocessing, attribute reduction and so on. Decision rules of rough sets given the right or similar decision-making from the condition, does not exclude the uncertain knowledge, provides a new method based on uncertainty of spatial data mining [2].

In view of rough set and decision tree has a strong complementary nature, this paper introduces rough set theory and decision tree in data mining field, uses original survey

data of Nong'an topsoil fertility survey and obtains the inductive generalization of the data by analysis, takes clustering samples by study. And then uses the rough set attribute reduction and decision tree method to generate the combination of soil fertility level evaluation rules. This method achieves good results in practice.

2 The basic principles of evaluation based on rough sets and decision tree

K-means clustering method makes the class as much as a compact cluster, separates as much as possible between classes and meets the needs of the soil data classification. Data of cluster centers, can basically cover the entire sample space, and makes the data more typical.

C4.5 algorithm is the most widely used decision tree algorithm, uses adoption rate of information gain to select attributes. Calculate the corresponding information gain ratio, and then select the highest division gain rate as the property of the information gain ratio, can handle continuous numeric attributes. Treatment strategy of missing values for some attribute is assigned to its most common value corresponding training examples, another more complex strategy is to give a probability for each possible value. In the tree formation, in order to prevent the uncontrolled growth of the tree, C4.5 algorithm uses a post-pruning method. The method evolves from one called "rule post-pruning" method. This method uses the training sample set to estimate their own errors before and after pruning, to determine whether the real pruning [3].

Johnson rough set attribute reduction algorithm can reduce attributes to the frequency by property size and weight, access a collection of the most relevant, can calculate the reduction effectively. Specific process is: According to a given decision table to calculate the discernibility matrix. Taking the attributes of largest frequency and weight to reduction sets, and then removes the attributes from resolution function in the same time and return the reduction when the resolution matrix is empty [4].

Therefore, this paper selects the K-means clustering method clustering soil data, extracts optimization samples, then applies the C4.5 decision tree algorithm and Johnson rough sets attribute reduction algorithm method to assist the fertility levels.

In the case of large amounts of data, the choice of learning samples is also an important part. This article uses the incremental method selected study samples, using cluster analysis for data processing, removing duplicate sample data, effectively reduce the study sample space and ensure the study sample data is typical. In this article, in order to avoid decision-making performance degradation by using cluster extraction method of learning samples, the author introduce a mechanism to re-select the sample, use of data mining process to provide a typical data to deduce a large data.

3 Applications Based on Rough Sets and decision tree

3.1 Experimental Data Acquisition

The experimental data is from the Nong'an survey data, provided by Agricultural Technology Promotion Center. Uses ARCGIS9.2 software vector the 1:10 10000 Nong'an soil map, the basic block diagram farmland protection and land-use maps, forms the digital soil map layers, and then matches the fertility survey and quality evaluation of cultivated land base map in 2006 at Nong'an, builds the Nong'an spatial database of sampling points of cultivated land survey and quality assessment.

Spatial database includes five aspects these are profile and physical chemical properties, site conditions, weather factors, soil nutrients, soil management, including plot latitude and longitude, land area, soil moisture, administrative, light radiation, annual rainfall, soil drought resistance, soil erosion, soil texture, groundwater depth, irrigation, crop rotation suitability, terrain position, soil parent materials, humus layer thickness, salinity, soil ph values, the effective copper, an effective iron, slowly available potassium, available potassium, effective manganese, total nitrogen, available phosphorus, organic matter, cation, effective zinc and level total of 27 properties, 3153 polygon data. According to Nong'an farmland agricultural requirements, the soil is divided into 6 grades 1, 2, 3, 4, 5, 6.

3.2 Extract Based on Clustering

Clustering method is used to cluster analysis of experimental data, removing duplication of learning samples, analyzing affects that the number of samples to decision-making ability of the decision tree model.

The author does data sample based on the principle of gradual, uses K-means clustering algorithm to cluster the original data sample. First，The author selects k objects as initial cluster centers from n data objects, and then according to their distance with these cluster centers, the assign remaining pairs of other objects to their most similar clustering; calculates with each new cluster of the cluster center; repeats this process until the standard measure function begin until convergence. In the experiment, clustering sample size are 200, 400, 700, 1000, 1200, 1400, 1800.

With the results of the clustering, using ordinary decision tree model to do sample study, samples from the cluster center, through the establishment of decision tree evaluation model to determine the number of learning samples, if the model of high accuracy, tend to converge, then the sample meets the needs, or else need to re-select learning samples. Decision-tree evaluation model test results as shown in Table 1 and Figure 1.

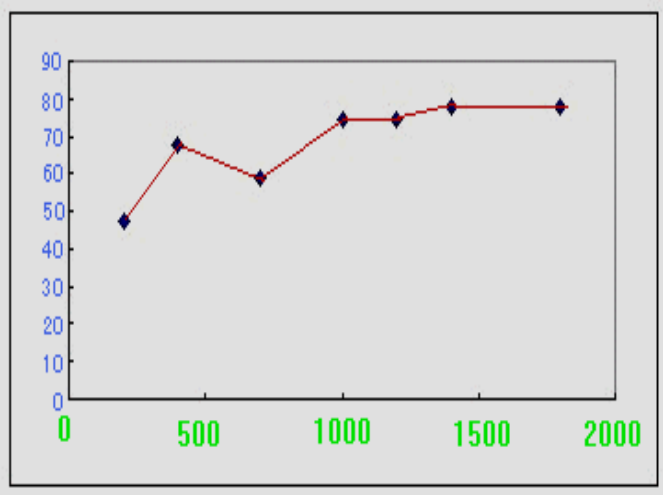

Fig.1. sample and model precision

Table 1. sample and model precision

Sample size (a)	200	400	700	1000	1200	1400	1800
Model accuracy（%）	47.32	67.85	58.93	74.10	74.10	77.68	77.68

From Table 1 and Figure 2 can see that, the curve has a clear inflection point in the 1000 samples, the accuracy of 1400 sample is improved 3.58 percentage points more than 1200 samples, but the 1800 sample evaluation model accuracy is the same evaluate the model with 1400 samples. From these data can see that on the basis of the 1400 samples for each increase of 200 samples, the model precision can be improved very little, indicating that in the 1400-based accuracy rates were stable, is the study the optimal number of samples.

3.3 Data Mining Based on Rough Set and Decision Tree combination

Algorithm for Mining Process. After optimization of the sample clustered a total of 1400 records, contains the soil area, soil moisture, administrative, light radiation, annual rainfall, soil drought resistance, soil erosion, the degree of soil texture, groundwater depth, irrigation, crop rotation suitability, terrain position, into the soil parent material, humus layer, salinity, soil ph values, the effective copper, effective iron, slowly available potassium, available potassium, effective manganese, total nitrogen, available phosphorus,

organic matter, cation, effective Zinc total 26 condition attributes and 1 decision attribute as soil fertility levels, the levels of Nong'an is divided into a total of six grades 1,2,3,4,5,6.

Rough set attribute reduction algorithm requires that data is discrete data, according to the soil data characteristics, the Entropy / mdl discrete algorithm is carried to do the data processing. Enter the 26 condition attributes to rough set algorithm and form a conditional attribute set C, the fertility level is as the decision attribute D. Using rough set reduction Johnson Reduction Method on attribute set C, get the simple property set. Its reduction properties are soil moisture, light radiation, annual rainfall, soil drought resistance, soil erosion, the degree of soil texture, groundwater depth, crop rotation suitability, terrain position, into the soil parent material, humus layer thickness, salinity, ph values, effective iron, slowly available potassium, available potassium, effective manganese, phosphorus, cation, a total of 19 condition attributes, and removed 7 redundant attributes.

After reduction, import the decision tree algorithm C4.5 to do decision-making. The data mining algorithm based on decision tree combining of rough sets are described below:

Continuously remove more important attributes from the condition attributes C relative to the decision attribute D, which makes the dependence degree D to it equal the dependence degree D to C, then obtains the attribute reduction set. Then, use the information gain as heuristic information, select attributes that can classify the sample well, building a branching, and divides the training set according the above, until there are no attributes which may divide again. Afterward use the test set to verify and modify the decision tree model. Algorithm flow shown in Figure 2

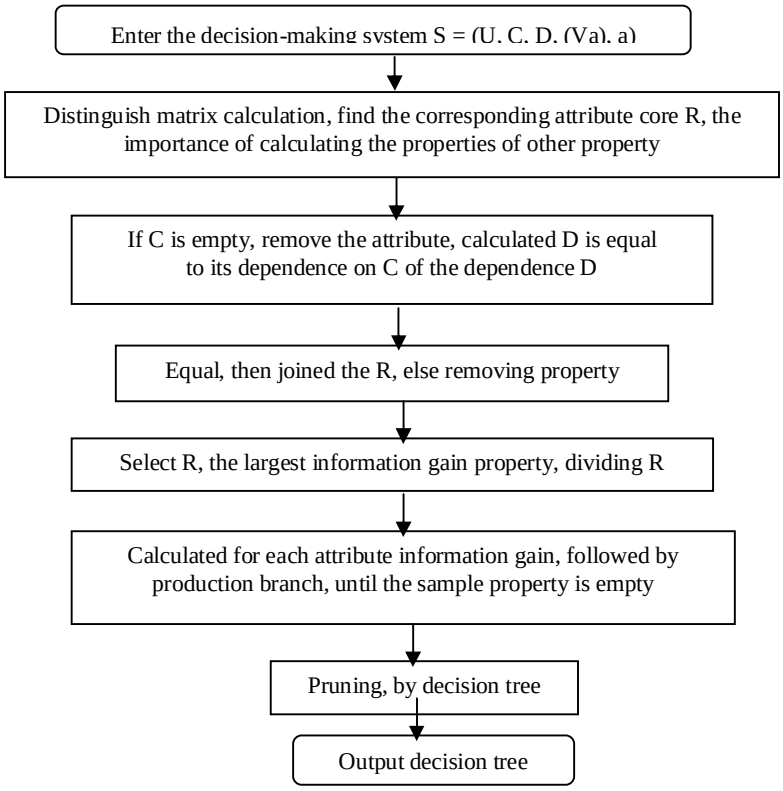

Fig.2. Algorithm flowchart

Data Mining Results. In 1400 Records, correctly classified data is 1298, the others are not. Decision tree model generated a total of 317 nodes, including leaf nodes 159. Generated part of the decision tree shown in Figure 3.

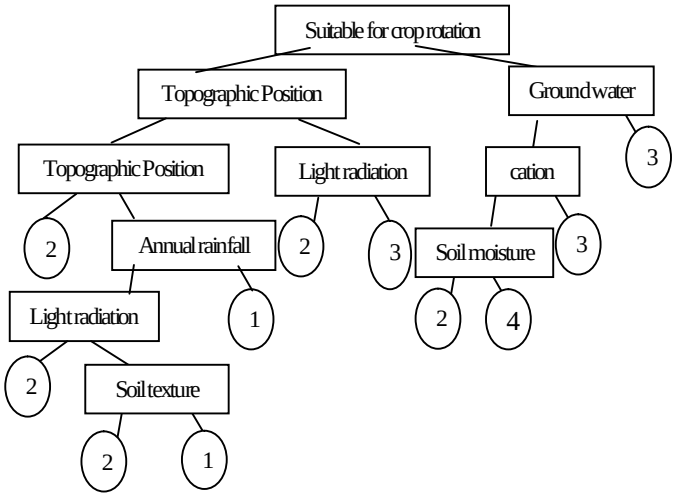

Fig.3. Part of the fertility level decision tree

According to the decision tree model, the author uses 100 records on the resulting decision-making model validation, validation results showes that: the decision-making accuracy rate of the decision-making model is 92% (Table 4.2), and extracted decision rules 159, some of the decision-making rules are as follows:

Rule 1: IF phosphorus> 21.775 AND salt content <= 1 AND light radiation intensity <= 2 AND soil drought resistance <= 3 AND phosphorus <= 22.3 AND soil moisture <= 22 THEN soil fertility grade = 1;

Rule 2: IF phosphorus> 21.775 AND salt content <= 1 AND light radiation intensity <= 2 AND soil drought resistance <= 3 AND phosphorus <= 22.3 AND soil moisture> 22 THEN soil fertility grade = 3;

Rule 3: IF phosphorus> 21.775 AND salt content <= 1 AND light radiation intensity <= 2 AND soil drought resistance <= 3 AND phosphorus> 22.3 AND Humus layer <= 3 AND topographic position <= 18 AND PH Value <= 7.25 THEN soil fertility grade = 1

To compare the feasibility and advance of this experimental methods, the author uses the same training set and test set, respectively run decision tree models tested that with no clustering, no rough set attribute reduction, clustering and attribute reduction and attribute reduction but no clustering, the comparison results as shown in table 2.

Table 2. Comparison of the table the results of several methods

Whether clustering	Whether reduction	The number of decision tree nodes	Number of leaf nodes	Model accuracy rate%	Removed redundant attribute
clustering	reduction	317	159	92	7
clustering	No reduction	329	165	90	0
No clustering	reduction	539	270	91	7
No clustering	No reduction	543	272	85	0

4 Rsults and Analysis

Experimental results show that:
(1).The samples using the clustering method taken, by the decision tree model checking, are more typical, can be used as a standard model of learning samples;
(2)Using the rough set attribute reduction and decision tree method of combining, reduce the tree's branches, removing 7 redundant attributes, the model accuracy rate is 92%, not only improves the mining efficiency, but also improve the accuracy of the model rate. Sun Weiwei's studies who evaluate the use of decision tree method for assessing soil quality and grade of the same data sets have shown that the decision tree method is better than exponential product method and gray opened a comprehensive evaluation method, often of the right index and France are very close[5]. The results of this study show that decision tree based on rough set model is better than an ordinary decision tree model.
(3)Comparison of the results from the correlation method can be seen: in accuracy, clustering methods are slightly higher than the non-clustering methods, reduction methods are slightly higher than the non-reduction methods, clustering and rough sets method combining are higher than all other methods. More important is the preferred method based on clustering of samples in addition to a large number of redundant samples, attributes reduction method based on rough set e excepts some redundant properties, saving time and space, reducing the complexity of the model.

In a word, this paper presents a combining method that clustering and rough set theory is better than other methods in time, space and accuracy, and achieved a good result in Land capability classification.

References

1.Xue Zhengpin, Deng Hua, Yang Weixing, et al. Based on decision tree and chart level superposition accurate agricultural output chart analysis method[J]. Agricultural engineering journal, 2006, 22(8):140-144.

2.Wu Chengdong, Xu Ke, Hang Zhonghua, et al. Decision tree's data mining method based on rough set. Northeast University journal (natural sciences version)[J],2006(05): 481-484.

3.Zhu Ming. Data Ming [M]. He Fei: Chinese Scientific and Technical University Publisher, 2002:67-76.

4.Z.Pawlak,Rough Sets.Internet Compute Inform Sci,1982,11(5):341-356.

5.Sun Weiwei, Hu Yueming, Liu Xingcai etc. For decisions the soil quality level research [J].z journal of south china agricultural university (Jcr science edition) 2005,7： 118-110.