Recognition of Pottery Culture Type in Southeast Heilongjiang Region Based on Fusion Rule Attribute Generalization Algorithm Based on Computer Software Calculation

Lei Zhao

Mudanjiang Normal University, Heilongjiang, 157000

*E-mail: 47872392@qq.com

Abstract. The high development of computer technology has brought new changes to all walks of life, and has also provided a new breakthrough for the development of some industries, which not only improves the quality, but also improves the efficiency. Combined with the new computer algorithm rules, the emergence of new computer software, to the future of the archaeological industry is a bleak scene. In this paper, the feasibility of the rule algorithm of computer software is confirmed by the recognition of pottery type in southeast Heilongjiang province.

Keywords: Computer Software, Fusion Rules, Attribute Generalization, Pottery Culture

1. Introduction
Ceramics is not only a necessity of life, but also a carrier of culture and art to some extent, which has played an important role in improving people's quality of life and aesthetic taste since ancient times. Ceramics in the southeast of Heilongjiang have profound cultural connotation and distinctive regional characteristics. Northeast culture has unique cultural characteristics and ideological connotation, combined with its excellent regional cultural characteristics, and it is optimized and inherited, applied to the design of ceramic products, forming ceramic products with distinctive cultural characteristics, and then maintaining a relatively stable continuous development[1-3]. It is of great significance to inherit and carry forward the treasures of Chinese traditional culture, promote the construction of enterprise brands, and spread the cultural ideas of characteristics.

Inductive learning is a very important data mining method. But because the amount of data in the database is often very large, which affects the efficiency of induction. It is necessary to take effective measures to reduce the data. Attribute-oriented data generalization and inductive learning techniques
are effective ways to solve this problem\cite{4}. Conceptual hierarchy is the basis of all data mining, and the attribute-oriented method can be used in relational database data mining after improvement and expansion, which is called attribute-oriented generalization method. This method combines machine learning, especially example learning technology, with set-oriented database operation, which greatly improves the function and efficiency of data mining. This paper uses the attribute data of the first stage culture of site, after necessary cleaning and pretreatment, carries on a series of generalization rule mining, and carries on the archaeological explanation.

2. Strategies for fusion rule attribute generalization algorithm

The progress of knowledge discovery research has promoted the rapid development of induction-based knowledge discovery methods, resulting in an attribute-oriented induction method, which generalizes low concept layer data to high level concept layer according to concept level, so that multi-layer or high-level rules can be found. This algorithm can combine the example learning method in machine learning with the operation technology of database, and convert the concept hierarchy tree corresponding to the attribute to the concept layer through the original data set generalization form, which can effectively reduce the size of the data set and reduce the computational complexity of the knowledge discovery process significantly.

2.1. Generalization on minimum decomposition units

Since the minimum decomposition unit of a dataset is a single attribute, generalization on the minimum decomposition unit is more convenient to see the generalization of generalization to avoid over statement. It is usually done on a single attribute of the dataset.

2.2. Remove irrelevant attributes

If an attribute in the relevant dataset has different values and it does not have a higher concept layer on the concept hierarchy tree, the attribute cannot be generalized to a higher concept layer, so the attribute can be removed from the discovery task.

2.3. Concept tree climb

If the attribute has a higher level in the concept hierarchy tree, the generalized data can centralize all records with that attribute value with a high level attribute value replacement.

2.4. Add CNT properties

As a result of the concept tree climb, there are usually many records with the same attribute values that can be merged into one. As a result, by adding CNT attributes to record the same records in the generalization table and accumulating the number of different records, it plays an important role in the process of knowledge discovery.

2.5. Threshold set

The rule correct rate is defined by using CNT attribute removal strategy P, and the trade-off rule threshold is defined L, indicating that the rule is valid when P>L and invalid when P<L. The purpose of any attribute climbing several times along its concept layer is to generalize the data set to a certain level, so that the inductive threshold can be set to control the process. If the attribute value reaches this threshold, the generalization does not need to continue, otherwise it needs to continue. The inductive threshold of the generalization table is to be set. If the number of records of the generalization table is greater than the threshold, it is necessary to continue the generalization until this inductive threshold is
satisfied.

2.6. Basic generalization algorithm description algorithm
The algorithm input condition has relational data set, learning task, attribute concept level, inductive threshold t[i], generalization table inductive threshold and so on. The output condition is the user expected generalization after data set. The algorithm consists of two steps:

First, according to the submitted learning task, the original relational data set is used to collect the attribute data to form the initial generalization set; then, the basic generalization algorithm is executed to generate the generalized data set.

3. Principle of the algorithm for the generalization of rule attributes

3.1. Methodological improvements
In the original algorithm, only the quasi-identifier attributes with the most values in the data table are generalized. In the face of having multiple quasi-identifiers with the same number of values, the algorithm fails. Then you need to process the data table before generalization. When there is only one quasi-identifier attribute with the most value, the algorithm directly selects this attribute for generalization, and then includes the number of values after generalization. When there are multiple quasi-identifiers with the same number of values, the concept of information entropy is used to calculate it, and the optimal generalization object is obtained. Information entropy is a measure of uncertainty of information\(^5\). Variable can be represented by the following formula.

\[
H(X) = -\sum_{i=1}^{n}P(x_i)\log_2P(x_i)
\]  

(1)

Where X represents a random variable, corresponding to a set of all possible outputs, defined as a set of symbols, and the output of a random variable is represented by X. P (x) represents the output probability function.

In this algorithm, when the entropy calculation is carried out on the attribute range of the identity, the higher the entropy value, the higher the data disorder, the less vulnerable to background knowledge attack; on the contrary, the smaller the entropy value, the smaller the attribute disorder, the more vulnerable to background knowledge attack. So the quasi-identifier attribute with smaller entropy value is preferred for generalization. One set of quasi-identifiers. For example, \{20345,20345,20345,20587,20618,20980\}, 20345 appears with a frequency of 3, the other three numbers appear with a frequency =1, the number of values is 4, then P(20345)=3/6=1/2, P(20587)=P(20618)=P(20980)=1/6, entropy is 1.793, and the other set of quasi-identifiers with a value of 4 \{12344,12356,12457,14390\}, P(12344)=1/4, P(12356)=1/4, P(12390)=1/4, entropy 2. So the generalization selects the object as the first collection.

3.2. Dynamic anonymity test
In order to avoid overgeneralization of a single attribute, it is necessary to test the anonymity of the dataset after each generalization. When the test result is "satisfied ", the system outputs the result, but when the test result is " not satisfied ", the system returns the quasi-identifier selection stage. In this process will also determine whether there are multiple values of the same maximum attribute, only one directly generalized, otherwise, into the entropy value calculation link. And so on, the final
generalization result is obtained after multiple cycles.

3.3. Algorithm thinking

First, enter a data table that needs to be generalized PT, need to preprocess the data table, delete sensitive attributes, extract a collection of quasi-identifiers, enter a value K generalization level, and enter a domain induction level tree. Then K-anonymous detection of the PT, when the quasi-identifier properties meet the requirements of K-anonymity, directly output table PT. When the table PT does not meet the requirements, the following processes are carried out: (1) statistics of the number of quasi-identifier attribute values and their range; (2) selection of the most valued set of quasi-identifier attributes to determine whether it is unique or not, if unique, to generalize, otherwise, multiple attribute sets with the largest number of values are entropy calculated, and the attributes with the smallest entropy value are selected for generalization; (3) judgment of whether the PT meets the requirement of K-anonymity after generalization, the output of the result after generalization, if not satisfied, the quasi-identifier selection, generalization and K-anonymity test continue; (4) Outputs PT.

of generalizations that satisfy K-anonymous tests The operation and display on computer software are as follows:

Input: Data table PT, set of quasi-identifiers, generalization level K value

Output: Table PT satisfying K-anonymity

1. n ← 0

2. if (K_anonymity (PTn)) // Judging whether the PT is satisfied K-anonymous

3. return PTn

4. else v = Statistic(PTn)// Statistics the number of attribute values of each quasi-identifier

5. x = Get.most(v)// Unsatisfied, select the set of quasi-identifier properties with the most values

6. if(x.sum > 1)// Determines whether there are multiple quasi-identifier attributes with the most values

7. i = cacShannonEnt(n.data) // Calculate the entropy of the property set with the maximum number of values

8. Gen(Get.min(i)) // Select the least entropy attribute for generalization

9. n++;

10. return 2; // Return to Step 2

11. else Gen(x)

12. return 2;

13. end. // End of run
4. Grouping and sequential data mining of houses

The housing situation on the whole site database and the ruins distribution map, and it is shown in the form of the ruins distribution map.

4.1. The whole grouping and sub-age data mining of ceramics

According to the concept hierarchy tree of ceramic grouping and culture layer relative layer table, the data are generalized, and the constraint conditions that must appear on the relic distribution map are satisfied, the quantity rule is produced, and the ceramic grouping and epoch are visualized. It can be seen from Figure1 that the number of ceramics in the early stage is not large, except for the characteristics of grouping. There are no other rules for the distribution of ceramics in each group. The medium term is different, with a large number of ceramics, which occupy most of the area of each group, and other large ceramic industries except super large ceramics appear in this period. The number of late ceramics is very small, except for a large ceramic, mostly small and practical ceramics.

![Figure 1](image_url)

(a) (b) (c)

**Figure 1.** Housing sub-age generated by attribute generalization based on computer software fusion rules

4.2. Packet sequence data mining for ceramics

The data are generalized according to the concept hierarchy tree established by ceramic grouping and culture layer relative horizon table, and the constraints that must appear on the relic distribution map are satisfied. Take the distribution of ceramics in each era in Figure2 as an example to illustrate as follows.

![Figure 2](image_url)

(a) (b) (c)
Figure 2. Housing sub-age based on generalization of computer software fusion rule attributes

There were only three ceramics in this group, three of which were practical ceramics. As the population grew, the group's mid-term residents lived separately in various parts of the group's residential areas, creating not only a large number of houses here, but also a significant increase in the amount of ceramic unearthed[6]. Enough to show that pottery occupies a very important position in the life of ancient residents. In addition, the emergence of ornamental pottery shows that social productivity has changed, people's spiritual and cultural life has also changed, pottery technology progress, not only increased production, but also more exquisite.

5. Conclusion
In this paper, the concept hierarchy tree is established by using computer software to calculate the metadata database of attributes and image data, the unearthed pottery and the relative layer table. By using the concept hierarchy tree and the attribute-oriented generalization algorithm based on the fusion rule, the attribute generalization of indoor unearthed pottery in a house in the southeast of Heilongjiang is carried out, and the number rule of the attribute generalization of the fusion rule is obtained. At last, the data mining of the whole house is carried out, and it is displayed in the form of visualization to illustrate the evolution of indoor ceramics. This method is helpful to improve work efficiency and provide scientific basis for the study of ceramic culture and archaeology.

Acknowledgments
The papers’ project is Study on Pottery Culture Device Shape in Southeast Heilongjiang Province.

References
[1] Sun Yi-ching, Bi Shuoben, Huang Jia-zhu, et al. Application of rule-based attribute-oriented generalization algorithm in settlement archaeology --- A case study of the first phase culture of Jiangzhai site [J]. Computer Engineering and Applications, 2005.

[2] Huang Jianguo. A Data Generalization Algorithm Based on Data Cube [J]. Computer and Information Technology, 2006, 000(007): P.44-47, 78.

[3] Fu Qi, Yi Chao. Research on Multi-attribute Generalization Technology Based on Datafly Algorithm [J]. Information and Computers (Theory Edition), No.405(11): 60-62, 2018.

[4] Wu Wenzhuo. A Brief Discussion on the Art of Pottery Literary Life Wenhai Art Garden, 2017, 000(011):35.

[5] Cai Yi. Identification of authenticity of Chinese ceramics [M]. Liaoning People's Publishing House, 2004.

[6] Yang Simeng. A Study on Identification and Design of Ceramic Products Based on Huxiang Culture [J]. Science and Education Guide: Electronic Edition, 2013.