Modern Applied Science
www.ccsenet.org/journal.html
Vol. 3, No. 8
August 2009

Data Mining Based on Fuzzy Rough Set Theory and Its Application in the Glass Identification

Ruying Sun
College of Information, Linyi Normal University, Linyi 276005, China
E-mail: srysd@163.com

Abstract

To overcome the disadvantage of determining artificially the class number, fuzzy C means clustering is introduced to fuzzify the continual attribute, and the best minute class number is obtained by cluster validity analysis. The relationship of glass composition and its application is excavated using data mining method in this paper.

Keywords: Data mining, Fuzzy clustering, Fuzzy rough Set, Glass identification

1. Introduction

Data Mining is defined as a large number of incomplete, noisy, fuzzy and random data extracted implicit in which people do not know in advance, but potentially useful information and knowledge, such as concepts, knowledge rules, restrictions, laws and so on. Rough set theory has been successfully applied to relational database data mining, such as A Fuzzy Search Method for Rough Sets in Data Mining, Mining Stock Price using Fuzzy Rough Set System and so on (Osei Adjei and LiChen , 2001, 980).

The continuous attributes must be discrete before extracting the rules using rough set theory. This process will result in some degree of information loss because discrete attribute values will not be retained in property values in the actual existence of numerical differences. French scholar D.Dubios and H.Prad presented the definition of fuzzy rough set combined rough set and fuzzy set to solve the problem of information loss in the course of attribute discrete based on rough set. Using the fuzzy rough set theory to deal with data sets can retain more original data set contains information.

However, the current attribute fuzzification method need to artificially divide into several classes, almost not considering the specific characteristics of attribute values. Methods are often too subjective, unreasonable and poor operability.

2. Fuzzy Clustering Analysis

2.1Fuzzy C-means Clustering Method (FCM)

$X=\{x_1,x_2,\cdots x_n\}\subset R^p$ is a limited data set in the feature space, n is the number of data items , c is the number of clusters with $2\le c\le n$, $R^{c\times n}$ is all the real matrix collection, $V=\{v_1,v_2,\cdots v_c\}\subset R^p$ is the vector collection in the feature space R^p, has c cluster center vector. $\mu_i(x_j)$ simply recorded as μ_{ij}, is the membership which is No. samples j belongs to No. centers i. The objective function of fuzzy C-mean clustering analysis is defined as follows.

$$\min J_m(U,V;X)=\sum_{i=1}^{c}\sum_{j=1}^{n}\mu_{ij}{}^m \| x_j - v_i \|_A^2$$

1) $\mu_{ij}\in[0,1],1\le i\le c,1\le j\le n$;

2) $\sum_{i=1}^{c}\mu_{ij}=1,1\le j\le n$;

3) $0<\sum_{j=1}^{n}\mu_{ij}<n,1\le i\le c$.

where

$$v_i = \frac{\sum_{j=1}^{n} u_{ij}{}^{m} x_j}{\sum_{j=1}^{n} u_{ij}{}^{m}}, 1 \le i \le c \tag{1}$$

$$\mu_{ij} = \frac{1}{\sum_{k=1}^{c} \left(\frac{d_{ij}}{d_{kj}} \right)^{1/m-1}} \tag{2}$$

$d_{ij} = \| x_j - v_i \|_A^2$, $1 \le i \le c, 1 \le j \le n$, is a distance measure between object x_j and cluster centre v_i, m is a weighting exponent on each fuzzy membership. A solution of the object function can be obtained via a process of iterating (1) and (2). The result is an optimal fuzzy division of x, $U^* = [\mu_{ij}{}^*]$.

2.2 Cluster Validity Analysis

Formula (3) defines the function of clustering validity.

$$FP(U;c) = \frac{1}{n}\sum_{j=1}^{n}(\sum_{i=1}^{c}\mu_{ij}^2 / \sum_{i=1}^{c}\mu_{ij}) - \frac{1}{c}\sum_{i=1}^{c}(\sum_{j=1}^{n}\mu_{ij}^2 / \sum_{j=1}^{n}\mu_{ij}) \tag{3}$$

Where c is the number of clusters, U is the membership matrix. If there is (U^*, c^*) to meet $FP(U^*, c^*) = \min_{c}\{\min_{\Omega_c} FP(U;c)\}$,then (U^*, c^*) is the optimal effective clustering, c^* is the number of the best classification.(GAO Xinbo, 2004, 59).

3. Data Mining Based on Fuzzy Rough Set

3.1 Attribute fuzzy-dependent analysis

Each equivalence class is fuzzy in the fuzzy-rough set, its lower and upper approximation is as follows. (Hongli Liang, Huaguang Zhang and Derong Liu, 2004, 584)

$$\mu_{\underline{X}}(F_i) = \inf_x \max\{1 - \mu_{F_i}(x), \mu_X(x)\}, \forall i$$
$$\mu_{\overline{X}}(F_i) = \sup_x \min\{\mu_{F_i}(x), \mu_X(x)\}, \forall i \tag{4}$$

Where, F_i is a fuzzy equivalence class, $\mu_X(x)$ is the membership which x belongs to arbitrary fuzzy set X in the domain U.

Formula (5) defines the fuzzy domain of F_i.

$$\mu_{POS_C}(F_i) = \sup_{X \in U|D} \mu_{\underline{X}}(F_i) \tag{5}$$

Where $F_i \in U \mid C$, X is one fuzzy equivalence class of decision attribute D.

$\mu_{POS_C}(x)$ is the degree of x ($x \in U$) belonging to the fuzzy domain.

$$\mu_{POS_C}(x) = \sup_{F_i \in U|C} \min\{\mu_{F_i}(x), \mu_{POS_C}(F_i)\} \tag{6}$$

According to the definition of fuzzy domain, the degree of decision attribute set D dependence on condition attribute set C can be obtained based on the fuzzy-rough set.

$$\gamma_C(D) = \frac{|\mu_{POS_C}(x)|}{|U|} = \frac{\sum_{x \in U} \mu_{POS_C}(x)}{|U|} \tag{7}$$

3.2 Attribute Reduction Algorithm

Attribute reduction is defined as deleting redundant attribute in the premise of maintaining the classification of decision table or decision-making ability. It can be expressed as the following definition (Richard Jensen and Qiang Shen, 2002, 30).

Definition1: If there is $C' \in C$ to meet

1) $\gamma_C(D) = \gamma_{C'}(D)$;

2) $\gamma_{C'-\{a\}}(D) < \gamma_{C'}(D)$, $\forall a \in C^{'}$

then C' is a reduction C compared to D. Where C is the condition attribute set and D is decision attribute set.

In order to reduce computational complexity, attribute reduction algorithm which attribute gradually reduced is used in this article. This method does not require verifying the conditions of each subset of attributes. Attributes that will not result in the loss of information of decision table system are gradually reduced from condition attributes set. Algorithm is as follows.

```
$R \leftarrow C$;
do;
$S \leftarrow \{\}$;
$\forall x \in R$;
$if\ \gamma_{R-\{x\}}(D) = \gamma_R(D)$;
$S \leftarrow S \cup \{x\}$;
$if\ \ s = \{\}$;
return R ;
$\gamma_{\circ} \leftarrow \gamma_C(D); T \leftarrow \{\}$;
$\forall x \in S$;
$if\ \gamma_x(D) < \gamma_{\circ}$;
$T \leftarrow \{x\}$;
$\gamma_{\circ} \leftarrow \gamma_x(D)$;
$R \leftarrow R - T$.
```

3.3 Data Mining Based on Fuzzy Rough Set Theory

The steps extracting rules from database are as follows.

First. Pretreatment. Complete loss of data, and delete the duplicate object, then structure the decision table.

Second. Data Reduction. Eliminate redundant attributes used attribute reduction algorithm based on fuzzy rough set and structure new decision table, and then eliminate redundant attribute value of the new decision table. Thus minimum reduction of decision table is obtained.

Third. Extract rules. Extract valuable rules according to the minimum reduction getting ahead.

4. Examples Analysis

Example 1

A simple data set of weather information.

First. Pretreatment.

Structure the decision table such as table 1.

The decision table contains a conditions attribute set C and a decision attribute d .where C={a1, a2, a3, a4}, a1 is the sunny index , a2 is the temperature index,a3 is the humidity index and a4 is the wind conditions index. Decision attribute d contains two values, d=0 indicates that he did not go out to play, d=1 play out.

Second. Data Reduction.

The fuzzy partitions matrix U of the four attributes are obtained by Fuzzy C-means clustering respectively. The results of clustering validity analysis are shown in table 2. From Table2 we can see, it is the most effective that a1, a2 are divided into 3 categories and a3, a4 2ategories. At the same time the partition of domain based on attributes are obtained as follows.

U/a1={{1,2,8,9,11},{4,5,6,10,14},{3,7,12,13}}

U/a2={{1,2,3,5,13},{6,7,9},{4,8,10,11,12,14}}

U/a3={{1,2,3,4,8,10,12,14},{5,6,7,9,11,13}}

U/a4={{2,4,6,7,9,10,11,12,14},{1,3,5,8,13}}

Attribute a3 can be reduced by the attribute reduction algorithm in this article. Eliminate redundant attribute value of the new decision table, then obtain a minimum reduction of decision table which show in table 3.

Third. Extract Rules.

The valuable decision rules are extracted as follows.

1) if $0.75 \le a1 \le 0.95$ and $0.85 \le a2 \le 0.95$ then d=0;

2) if $0.45 \le a1 \le 0.6$ then d=1;

3) if $0.45 \le a2 \le 0.7$ and $0.4 \le a4 \le 0.65$ then d=1;

4) if $0.1 \le a1 \le 0.25$ and $0.4 \le a4 \le 0.65$ then d=0;

5) if $0.75 \le a1 \le 0.95$ and $0 \le a4 \le 0.3$ then d=0;

6) if $0.75 \le a1 \le 0.95$ and $0.2 \le a2 \le 0.4$ then d=1.

The results showed that the practical significance and a smaller set of rule can be explored by the algorithm used in this paper.

Example 2

Glass Identification Database (http://ftp.ics.uci.edu/pub/machine-learning-databases/glass).

Glass Identification Database contains 214 instances, each instance include 9 attributes and a decision attribute. C={c1,c2,…c9}={RI, Na, Mg, Al, Si, K, Ca, Ba, Fe}, all attributes are continuously valued. where, c1(RI) is refractive index, unit measurement of c2 to c9 is weight percent in corresponding oxide. Decision attribute d is the type of glass, it has 7 discrete values.

1--building_windows_float_processed;

2--building_windows_non_float_processed;

3--vehicle_windows_float_processed;

4--vehicle_windows_non_float_processed ;

5--containers;

6--tableware;

7--headlamps.

Attribute c1, c3, c6, c9 can be reduced by the attribute reduction algorithm in this article. Eliminate redundant attribute value of the new decision table, then obtain a minimum reduction of decision table, and 31 valuable decision rules are extracted.

The results showed that 31 useful decision rules are excavated from the original 214 information indicates used the proposed data mining algorithm, the 183 redundant information indicates are deleted, simplifying the basis for judging. And the 31 rules show the relationship between glass compositions and their application. As long as the weight percent of Na, Al, Si, Ca, Ba in its corresponding oxide can be judged that the use of glass is confirmed. In other words, for an unknown glass, as long as the content of Na, Al, Si, Ca, Ba in its corresponding oxide is detected, the source of glass can be determined, thus the identification methods for glass is simplified.

5. Conclusions

Data mining based on fuzzy rough set provides an effective way to resolve the continuous attributes database mining. Limited useful information can be excavated from massive data by this method. It has great significance of saving the data storage and reducing the explosion possibility of information systems. Also it has important practical significance in the production and related projects.

References

Gao, Xinbo. (2004). *Fuzzy Cluster Analysis and its Application*, Xidian University Press (chapter5).

Hongli, Liang, Huaguang, Zhang, and Derong, Liu. (2004). Roughness of Fuzzy Sets Based on Two New Operators. *0-7803-8353-2/04©2004 IEEE*, 583-586.

http://ftp.ics.uci.edu/pub/machine-learning-databases/glass.

Osei, Adjei and LiChen. (2001). A Fuzzy Search Method for Rough Sets in Data Mining. *0-7803-7078-3/01©2001 IEEE*, 980-985.

Richard Jensen and Qiang, Shen. (2002). Fuzzy-Rough Sets for Descriptive Dimensionality Reduction. *0-7803-7280-8/02©2002 IEEE*, 29-34.

Table 1. Decision table of weather information

No	a1	a2	a3	a4	d
1	0.8	0.9	0.9	0.3	0
2	0.75	0.85	0.88	0.5	0
3	0.5	0.95	0.75	0.2	1
4	0.2	0.6	0.8	0.4	1
5	0.15	0.95	0.5	0	0
6	0.25	0.3	0.55	0.6	0
7	0.45	0.2	0.55	0.65	1
8	0.78	0.7	0.85	0.1	0
9	0.9	0.4	0.45	0.4	1
10	0.2	0.65	0.9	0.45	1
11	0.95	0.55	0.6	0.55	1
12	0.5	0.55	0.8	0.6	1
13	0.6	0.9	0.6	0.2	1
14	0.1	0.45	0.85	0.5	0

Experimental steps are as follows.

Table 2. The results of clustering validity analysis for weather information

Class	FP(U; c)			
Number	a1	a2	a3	a4
2	0.0358	0.0974	-0.0034	-0.0369
3	-0.0069	0.0334	0.0084	-0.0241
4	0.0156	0.0434	0.0028	0.0356
5	0.0073	0.0663	0.0052	-0.0182

Table 3. A minimum reduction of decision table

No	a1	a2	a4	d
1	1	1	×	0
2	3	×	×	1
3	×	3	1	1
4	2	×	1	0
5	1	×	2	0
6	1	2	×	1