International Journal of Security and Its Applications
Vol.8, No.5 (2014), pp.195-202
http://dx.doi.org/10.14257/ijsia.2014.8.5.18

Data Recovery Method for Seafood Quality Safety System Based on Rough Set Theory

Xu E, Shuang Lin and Lulu Jin

Department of Information Science & Technology, Bohai University, Jinzhou 121000, China

Abstract

To solve the seafood quality and safety information table missing data problem, a method for filling missing data based on rough set was proposed. This method first determine whether there is missing data in information table, if there is, then the information table is divided into two parts, one for the complete information table, and the other part is totally incomplete information table, and then complete the information sheet structure similar matrix to calculate the attributes importance, then calculate the number of missing attributes in incomplete information table, and the calculated results conjunctive operation to fill the missing data. If it is not, then output directly. Numerical examples show that the method can be used filling the missing data in seafood quality and safety information table effectively.

Keywords: *Rough sets, seafood, fill, similarity matrix, quality and safety, frequency function.*

1. Introduction

Seafood quality and safety is not only related to our daily life and physical health closely, but also related to the economic development of China's exports of seafood. The seafood quality and safety evaluation of monitoring data needed to detect, constitute information systems and analyzes the data to obtain quality and safety. However, in practical applications, information systems are often due to technical limitations and data loss problems caused by lack of information, to data analysis and to establish evaluation index system difficult, not easy to obtain scientific and accurate seafood quality and safety conditions, and therefore on seafood quality and safety to fill in missing data is particularly important [1]. When there is information missing information systems, rough set theory can be used to solve. Rough set theory is an extension of set theory, the ability to classify data based study, presented by the Polish mathematician Z. Pawlak in the early 1980s, rough set theory can be the case in the absence of data on prior knowledge, analysis and vague and uncertain data processing is uncertainty mathematical research tool for studying imperfect and incomplete information described in knowledge acquisition [2]. So the rough set theory is used to fill the seafood quality and safety data, it can reduce the difficulty caused by the missing data to evaluate the quality and safety of seafood brought.

This paper defines a seafood quality and safety data filling algorithm based on rough set theory. The algorithm is not the best, but it can help to reduce complexity.

ISSN: 1738-9976 IJSIA

2. Related Definitions and Theorems

2.1. Related Definitions

To solve the problem of seafood quality and safety data filling in this study, related definitions and theorems referenced as follows [3-11].

Definition 1. Information System In rough set theory, an information table knowledge representation system $S=<U, A, V, f>$, where:

$U=\{x_1,x_2,\ldots,x_{|U|}\}$ represents domain;

$A= C\cup D$ represents a finite non-empty set of attributes, a subset $C=\{a_1,a_2,\ldots, a_i\}(i=1,2, \ldots,n)$ represents the condition attribute set, $D=\{d\}$ represents the result set of attributes;

f: $U\times A\rightarrow V$ represents information function .

Definition 2. Incomplete Information System In the information system $S=<U, A, V, f>$, if the attribute value V_a exist omissions or missing, the missing attribute values recorded as *, that is, in information system $S=<U, A, V, f>$, $\exists\forall\ V_a=*$; claimed that the information system S is incomplete information system.

Definition 3. Indiscernibility Relation Given a information table expression system $S=<U, A, V, f>$, for each subset of attributes $B\subseteq A$, definitions are indiscernibility relation $IND(B)$, that is $IND(B) =\{(x, y)|(x, y)\in U^2, V_b\in B\ (b(x)=b(y))\}$.

Definition 4. For each subset $X\subseteq U$ and indiscernibility relation B, the upper approximation and lower approximation of X are defined respectively by the basic set B as follows:

Upper approximation: $\bar{B}(X)=\cup\{x\in U|\ IND(B)\wedge I(x)\cap X\neq\varnothing)\}$;

Lower approximation: $\underline{B}(X)=\cup\{x\in U|\ IND(B)\wedge I(x)\subseteq X)\}$;

Positive domain: $Pos_b(X)=\underline{B}(X)$.

Definition 5. Similarity Matrix Let be information system $S=<U, A, V, f>$, $a_i(x_j)$ represents the values of sample x_j on attribute a_i. M(i,j) represents the similarity matrix element in the i-th row j-th column, the similarity matrix M can be defined as follows:

$$\mathbf{M}(i, j)=\begin{cases}\{a_k\mid a_k\in C\wedge a_k(x_i)=a_k(x_j)\}, & d(x_i)\neq d(x_j)\\ 0, & d(x_i)=d(x_j)\end{cases}.$$

Definition 6. Attribute Importance For F is the classification derived from attribute set B, the importance of attribute subset B' in the attribute set B is defined as $r_B(F)$ - $r_{B\setminus B'}(F)$; where $B'\subseteq B$, if attribute set B is default to the condition attribute universal set, then referred to as the importance of attribute subset B'; It can be also defined as $Pos_{B\setminus B'}(F)/\ Pos_B(F)$, where $Pos_B(F)=\cup_{X\in F}\ Pos_B(X)$.

Definition 7. Frequency Function In similarity matrix defines the frequency function $p(a_k)$, indicates the number of occurrences of attribute a_k in the similarity matrix, then the frequency function for the not same attributes can be defined as:

$$p'(a_k)=\sum_{\substack{i=1,j=1\\ i\neq j}}^{n}(C-\mathbf{M}(i, j))\,p(a_k).$$

Definition 8. In incomplete information system $S=<U, A, V, f>$, letbe $x_i\in U$, then can be defined:

$MAS_i=\{a_k|a_k(x_i)=*,k=1,\ldots,m\}$, represents loss attribute set of object x_i;

$MOS=\{x_i|MAS_i\neq\varnothing, i=1,\ldots,n\}$, represents loss object set of information system S;

$UNMOS=U/MOS$, represents non-lost object set;

AS_i={j| M(i,j)=U, $i \neq j$, j=1,…,n}, represents undifferentiated objects set of obje**2.2. Related Theorems**

Theorem 1. Let be incomplete information Table S=<U, A, V, f>, where A= $C \cup D$, $a_i \in C$, its complete and incomplete subset information Table respectively were S_1 and S_2, filling the redundancy missing data in S_2 , the classification ability of S does not change.

Prove: Let be a_i is the core attributes in S_1, after the removal of a_i, $\exists x, y \in U$, $\forall a_j \in C$ with $a_i(x)=a_j(y) \wedge d(x) \neq d(y)$, where $d \in D$, generates decision conflicts in S_1, and $\because U_1 \in U$, after the removal of a_i, S will also generate the same problem, a_i is still the core attributes in S, therefore, it prove that filling the redundant data cannot change the classification ability of information table .

Theorem 2. In similarity matrix, if there is any element M(i,j), its values for the set of all condition attributes, that all the conditions the same as the value of the attributes, then there is a conflict (inconsistent) information in the original information table S.

Prove: Assume that when M(i,j) values for all attributes collection criteria, there is no conflict (inconsistent) information in the original information table S, then, in the same condition attribute similarity matrix values of any element that M(i,j)={ a_k | $a_k \in C$, a_k (x_i)=a_k (x_j)}=C, $\because$ there is no conflict information in the original information Table S, there is $d(x_i)=d(x_j)$, at a time when the definition of the similarity matrix $d(x_i) \neq d(x_j)$ contradictory, so theorem proved.

Theorem 3. In similarity matrix, when Card(M(i,j))=Card(C)—1, C—M(i,j) belongs to the core attributes which Card(M(i,j) denote similar elements in the matrix M(i,j) contains a number of attributes.

Prove: In certain similarity matrix, undifferentiated set of attributes for the Card(C)—1 attributes, then the whole condition attribute set only one difference attribute that conditions in the whole set of attributes, only one attribute can distinguish these two objects, so C—M(i,j) belongs to the nuclear attributes.

Theorem 4. If a decision information system S=<U, A, V, f> (A= $C \cup D$) is consistent, there is $POSc\{d\}=U$.

Prove: Assume $POSc\{d\} \neq U$, when the object decision values are same, condition attributes of at least two objects are consistent completely, and the two attributes will be discernibility set. This is contradictory to the theory: In a decision information system, C is condition attributes and D is decision attribute. If the values of decision D of corresponding condition attribute of C are same in the indiscernibility, the decision system is consistent; Otherwise, it is incompatible. So $POSc\{d\}=U$.

3. The Basic Thought of Data Filling Algorithm

The basic thought of the data filling algorithm is first to determine whether the original data Table is complete information Table or not; If it is complete information Table, then output the information Table directly without processed; Otherwise, the information Table is divided into two parts, one part complete information is a subset of the complete information table, the other part is a subset of the incomplete information Table; Then structure the similar matrix based on a subset of the complete information Table and calculate attribute importance, while computing the number of missing data objects for each subset in the incomplete information Table, the results of conjunctive calculation; and then fill the missing attribute data; Finally, to get complete information Table. This can avoid the problems of processing the data of seafood quality and safety caused by missing information and inconsistent

information in the information systems.

The flow chart shown in Figure 1:

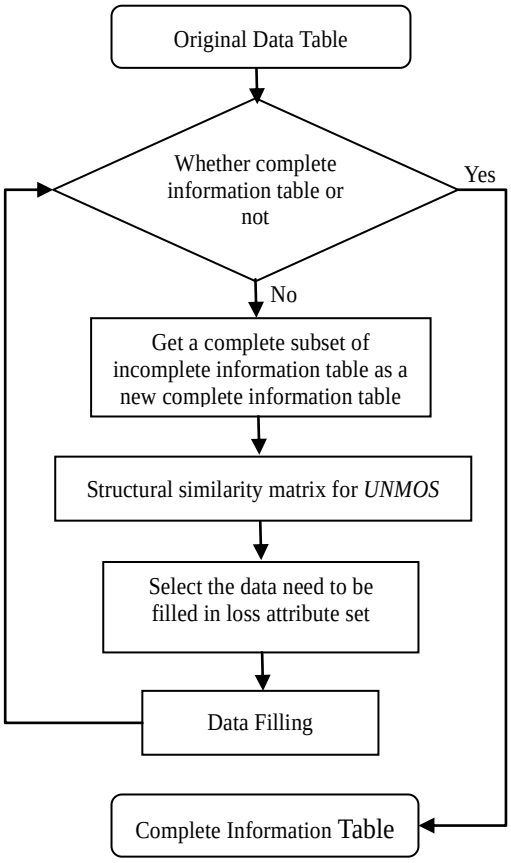

Figure 1. Seafood Quality and Safety of Data Filling Algorithm Flowchart

4. Data Filling Algorithm Study

Fill algorithm combines the basic ideas of data and processes, the basic steps of the algorithm are:

Input: Original information Table S_0=<*U*, *A*, *V*, *f*>;

Output: Padded data information Table.

Step 1: Inspection of the original data Table

if $\exists\forall V_a$=*, then (2);

else output;

Step 2: Calculate *MOS*、*UNMOS*、MAS_i, *IND*(*d*); *Q*=*IND*(*d*)∩*UNMOS*; S_1=<*U*, *A*, *V*, *f*>, *U*=*UNMOS*;

Step 3: Construct similarity matrix **M**(*i*,*j*)for new information Table S_1,

if (M(*i*,*j*)=*U*)∧$i \neq j$) (x_i, $x_j \in U$),

x_i=min{$q'(x_i)$, $q'(x_j)$}, *U*=*UNMOS*/ x_i;

else turn (3);

Step 4: Calculate the importance of attributes

if (*Q*=∅), $r_B(F) - r_{B\setminus B}(F)$;

else $a_k \in C$, calculate $p'(a_k)$, and let be $p'(a_k) = \max\{ p'(a_k)\}$;

Step 5: Select the attributes to be filled

Calculate $|MAS_i|$,

if $(a_k \in MAS_i \wedge max\{p'(a_k)\} \wedge min|MAS_i|)$,

$*= x_i(a_k)$;

else $*= x_j(a_k)$;

Step 6: Fill the missing data

while($MOS \neq \varnothing$) do

calculate $AS(x_i)$;

if ($AS(x_i)=y$), $x_i(a_k)= x_i(y)$;

if ($AS(x_i)=\{y_1,y_2,\ldots,y_m\}$), $x_i(a_k) = x_i(y_i)$.//where potential maximum of y_i equivalence classes, fill other attributes.

$UNMOS=UNMOS \cup \{x_i\}$;

Step 7: Output the new information Table.

5. Example

To illustrate this algorithm, the calculation process is exemplified. Let the class seafood seashells initial information table $S=<U, A, V, f>$, shown as Table 1, where $U=\{x_1,x_2,\ldots,x_{10}\}$, $A= C \cup D$, sub-set $C=\{a_1,a_2,\ldots,a_5\}$, $D=\{d\}$.

Condition attributes $a_1,a_2,\ldots,a_5$ sequentially represent sensory indicators (seafood appearance, texture, smell, *etc.*,), heavy metal pollution (mercury, arsenic, volatile basic nitrogen, BHC, DDT, *etc.*, (mg/kg)), fishing drug residues (antibiotics, hormones, chlortetracycline, oxytetracycline, tetracycline and chloramphenicol (μg/kg)), pathogens, microbial indicators (viruses, bacteria, total (cfu/g)) and toxins (toxins diarrhea, nervous toxins, *etc*.,); decision attribute *d* is used to represent security, general and unsafe, respectively, the corresponding index is 0,1,2; * represents missing data [8].

Table 1. Seafood Initial Information Table

U	a_1	a_2	a_3	a_4	a_5	D
x_1	7.4	0.7	34	3.51	0.56	0
x_2	7.8	0.88	67	3.2	0.67	0
x_3	7.8	0.76	54	3.26	0.65	0
x_4	11.2	0.28	61	3.16	0.55	1
x_5	7.4	*	34	3.51	0.56	0
x_6	7.4	0.66	40	*	0.56	0
x_7	7.9	0.6	59	3.3	0.46	0
x_8	7.3	0.65	21	3.39	0.47	2
x_9	*	0.58	18	3.36	0.57	2
x_{10}	7.5	0.5	102	3.35	0.8	0

According to the missing data filling algorithm for processing the initial information Table, first judge seafood initial information table in Table 1 whether incomplete information Table by definition 2; In Table 1, $\exists \forall\ V_a=*$, so the initial information Table is incomplete information system; Then need to fill missing data in incomplete initial information Table, so that the initial information table completeness; Then calculated in accordance with the definition 8 are:

$MOS=\{x_5, x_6, x_9\}$;

$UNMOS=\{x_1, x_2, x_3, x_4, x_7, x_8, x_{10}\}$;
$MASx_5=\{a_2\}$, $MASx_6=\{a_4\}$, $MASx_9=\{a_1\}$;
$U|IND(d=0)=\{x_1, x_2, x_3, x_5, x_6, x_7, x_{10}\}$;
$U|IND(d=1)=\{x_4\}$;
$U|IND(d=2)=\{x_8, x_9\}$;
$UNMOS \cap U|IND(d=0)=\{x_1, x_2, x_3, x_7, x_{10}\}$;
$UNMOS \cap U|IND(d=1)=\{x_4\}$;
$UNMOS \cap U|IND(d=2)=\{x_8, x_9\}$.

At this point there is complete information system S_1, where $U = UNMOS$, then complete information table shown in Table 2.

Table 2. Seafood Subset Complete Information Table

U	a_1	a_2	a_3	a_4	a_5	D
x_1	7.4	0.7	34	3.51	0.56	0
x_2	7.8	0.88	67	3.2	0.67	0
x_3	7.8	0.76	54	3.26	0.65	0
x_4	11.2	0.28	61	3.16	0.55	1
x_7	7.9	0.6	59	3.3	0.46	0
x_8	7.3	0.65	21	3.39	0.47	2
x_{10}	7.5	0.5	102	3.35	0.8	0

Then structure similarity matrix for *UNMOS* in Table 2 by definition 5 and the matrix shown in Figure 2.

$$\mathbf{M}(i,j)=\begin{pmatrix} U & 0 & 0 & \varnothing & 0 & \varnothing & 0 \\ & U & 0 & \varnothing & 0 & \varnothing & 0 \\ & & U & a_1 & 0 & \varnothing & 0 \\ & & & U & \varnothing & \varnothing & \varnothing \\ & & & & U & \varnothing & 0 \\ & & & & & U & \varnothing \\ & & & & & & U \end{pmatrix}$$

Figure 2. UNMOS Similarity Matrix

Then calculated attributes importance in the new information Table *UNMOS* according to the definition 6 and the results is: $a_1 < a_2 = a_3 = a_4 = a_5$; By $MASx_5=\{a_2\}$, $MASx_6=\{a_4\}$, $MASx_9=\{a_1\}$ known, $|MASx_5|=1$, $|MASx_6|=1$, $|MASx_9|=1$; This time, in accordance with the conditions $\{a_k \in MAS_i \wedge \max\{p(a_k)\} \wedge \min\{|MAS_i|\}$ select attribute a_k of the missing objects x_i be filled, that $x_5(a_2)$ in information Table; determined according to the algorithm $x_5(a_2)=0.7$; $UNMOS=UNMOS \cup \{x_5\}=\{x_1, x_2, x_3, x_4, x_5, x_7, x_8, x_{10}\}$; then structural similarity matrix for *UNMOS*, repeat the above steps, when $MOS=\varnothing$, access to new information Table as shown in Table 3.

Table 3. Seafood Complete Information Table

U	a_1	a_2	a_3	a_4	a_5	D
x_1	7.4	0.7	34	3.51	0.56	0
x_2	7.8	0.88	67	3.2	0.67	0
x_3	7.8	0.76	54	3.26	0.65	0
x_4	11.2	0.28	61	3.16	0.55	1
x_5	7.4	0.7	34	3.51	0.56	0
x_6	7.4	0.66	40	3.51	0.56	0
x_7	7.9	0.6	59	3.3	0.46	0
x_8	7.3	0.65	21	3.39	0.47	2
x_9	7.3	0.58	18	3.36	0.57	2
x_{10}	7.5	0.5	102	3.35	0.8	0

6. Conclusion

The algorithm can be proved by application examples that information Table with missing data were filled and the missing data problems of incomplete data and time span is too large in seafood quality and safety are solved. Meanwhile, the process of filling data, the algorithm used similarity matrix in constructing the matrix, to avoid the introduction of conflicting data, simplify data Table; and do not need to calculate the upper approximate and lower approximation when the best case, reduce the computational complexity of attribute importance. Compared with the delete method, the algorithm can handle small amount of information, there is a lack of information objects more cases; Solve missing data of seafood safety and quality issues. Filling data is the basis of the data processing, and the next step would be to fill the post-processing information table.

Reference

[1] C. Li and Y. Huang, "Food Safety Monitoring and Early Warning System", Chemical Industry Press, Beijing **(2006)**.
[2] Z. Pawak, "Rough Set", International Journal of Computing and Information Sciences, vol. 11, no. 1, **(1982),** pp. 341-350.
[3] W. Zhang, W. Wu, J. Liang and D. Li, "Rough Set Theory and Method", Science Press, Beijing **(2001)**.
[4] G. Wang, "Rough Set Theory and Knowledge Acquisition", Xi'an Jiaotong University Press, Xi'an, **(2001)**.
[5] Z. Shi, "Knowledge Discovery (2nd edition)", Tsinghua University Press, Beijing, **(2011)**.
[6] Q. Zhang, G. Wang and Xiao, "Approximation Sets of Rough Sets", Journal of Software, vol. 23, no. 7, **(2012),** pp. 1745-1758.
[7] G. Wang, Y. Yao and Yu, "A Survey on Rough Set Theory and Applications", Computer Journal, vol. 32, no. 7, **(2009),** pp. 1229-1246.
[8] X. E, X. Gao, S. Wu and Q. Zhang, "A New Method of Packing the Missing Data", Beijing University of Technology, vol. 27, no. 3, **(2005),** pp. 364-366.
[9] C. Jin, X. E, H. Mu and Y. Li, "Data Filling Method Based on New Relationship Matrix", Computer Engineering, vol. 37, no. 19, **(2011),** pp. 28-31.
[10] Q. Zhu, G. Zhang, T. Feng, P. Huang and D. Hou, "Study on Water Quality Analysis and Early-warning Technology based on Rough Set and Evidence Theory", Journal of Zhejiang University (Agriculture and Life Sciences), vol. 38, no. 6, **(2012),** pp. 747-754.
[11] J. Wu and H. Zou, "Attribute Reduction Algorithm based on Importance of Attribute Value", Computer Applications and Software, vol. 27, no. 2, **(2010),** pp. 255-257.
[12] S. Xiong, "Fresh Seafood Storage and Testing", Chemical Industry Press, Beijing, **(2007)**.