International Journal of Engineering, Mathematical and Physical Sciences
ISSN: 2517-9934
Vol:7, No:5, 2013

An application of the data mining methods with decision rule

Xun Ge and Jianhua Gong

Abstract—Rankings for output of Chinese main agricultural commodity in the world for 1978, 1980, 1990, 2000, 2006, 2007 and 2008 have been released in United Nations FAO Database. Unfortunately, where the ranking of output of Chinese cotton lint in the world for 2008 was missed. This paper uses sequential data mining methods with decision rules filling this gap. This new data mining method will be help to give a further improvement for United Nations FAO Database.

Keywords—Ranking, output of the main agricultural commodity, gross domestic product, decision table, information system; data mining, decision rule

I. Introduction

Recently, Food and Agriculture Organization of the United Nations has released rankings for output of Chinese main agricultural commodity in the world, i.e., the following data table was given ([10]), where $u_1, u_2, u_3, u_4, u_5, u_6, u_7$ denote 1978,1980,1990,2000,2006,2007,2008.

TABLE I: INCOMPLETE DATA TABLE

Item	u_1	u_2	u_3	u_4	u_5	u_6	u_7
Cereals	2	1	1	1	1	1	1
Meat	3	3	1	1	1	1	1
Cotton Lint	3	2	1	1	1	1	
Soybeans	3	3	3	4	4	4	4
Groundnuts in Shell	2	2	2	1	1	1	1
Sugar Cane	7	9	4	3	3	3	3
Tea	2	2	2	1	1	1	1
Fruit	9	10	4	1	1	1	1

The above data in Table 1 also appear in [1]. In Table 1, numbers in the first row denote years and others denote rankings. Unfortunately, Table 1 is an incomplete Data table, which does not give the ranking for output of Chinese cotton lint in the world for 2008. Thus, the following question arise naturally.

Question 1.1: Can we obtain the ranking of output of Chinese cotton lint in the world for 2008 from information hidden in Table 1?

As an agricultural country, it is a clear fact in China that output of the main agricultural commodity plays an important role in gross domestic product (abbr. GDP) If we take the set consisting of 1978, 1980, 1990, 2000, 2006, 2007 and 2008 as the universe of discourse, take cereals, meat, cotton lint, soybeans, groundnuts in shell, sugar cane, tea and fruit in Table 1 as condition attributes, take gross domestic product as the decision attribute, and take rankings as attribute values, then we can construct the following incomplete decision table with a missing attribute value, where rankings for Chinese gross domestic product in the world for 1978, 1980, 1990, 2000, 2006, 2007 and 2008 were given by United Nations Development Porgram in [9] (also see [2]).

TABLE II: INCOMPLETE DECISION TABLE I

Item	u_1	u_2	u_3	u_4	u_5	u_6	u_7
Cereals	2	1	1	1	1	1	1
Meat	3	3	1	1	1	1	1
Cotton Lint	3	2	1	1	1	1	
Soybeans	3	3	3	4	4	4	4
Groundnuts in Shell	2	2	2	1	1	1	1
Sugar Cane	7	9	4	3	3	3	3
Tea	2	2	2	1	1	1	1
Fruit	9	10	4	1	1	1	1
GDP	10	11	11	6	4	4	3

Thus, Question 1.1 is transformed to the following question, which makes it possible for us to use data mining methods solving Question 1.1.

Question 1.2: Can we mine the missing attribute value in Table 2?

In this paper, data mining methods with decision rule is introduced. We will use typical sequential data mining methods handling the missing attribute value in Table 2. For every possible value of the missing attribute value, we use decision rule checking the support, the strength, the certainty factor and the coverage factor of the decision rule. Consequently, we obtain a certain value of the missing attribute value. Results of this paper give an answer for Question 1.2 (hence for Question 1.1), which will be help to give a further improvement for United Nations FAO Database.

II. Values of the missing attribute value

In this section, we use sequential data mining methods giving all possible values of the missing attribute value "GAP" in Table 1. Sequential data mining methods were categorized by J. W. Grzymala-Busse in [3], which include mainly replacing a missing attribute value by the most common value of that attribute, replacing a missing attribute value by the mean for numerical attributes, assigning all possible values to the missing attribute value and assigning to a missing attribute value the corresponding value taken from the closest fit case.

X. Ge is with School of Mathematical Sciences, Soochow University, Suzhou 215006, China (e-mail: gexun@suda.edu.cn)
J. Gong is with Department of Mathematics, United Arab Emirates University, Al-Ain, United Arab Emirates (e-mail: j.gong@uaeu.ac.ae).

The following proposition give concrete explanations for the above methods ([3]).

Proposition 2.1: Sequential data mining methods include mainly the following.

(1) The most common value of an attribute: Every missing attribute value is replaced by the most common value of this attribute.

(2) The most common value of an attribute restricted to a concept: A modification of the method of replacing missing attribute values by the most common value is a method in which the most common value of the attribute restricted to the concept is used instead of the most common value for all cases, where a concept is the set of all cases with the same decision value.

(3) Assigning all possible attribute values to a missing attribute value: Every case with missing attribute values is replaced by the set of cases in which every missing attribute value is replaced by all possible known values.

(4) Assigning all possible attribute values restricted to a concept: Every case with missing attribute values is replaced by the set of cases in which every attribute a with the missing attribute value has its every possible known value restricted to the concept to which the case belongs.

(5) Replacing missing attribute values by the attribute mean: Every missing attribute value for a numerical attribute is replaced by the arithmetic mean of known attribute values.

(6) Replacing missing attribute values by the attribute mean restricted to a concept: Every missing attribute value of a numerical attribute is replaced by the arithmetic mean of all known values of the attribute restricted to the concept.

(7) Global closest fit: Replacing a missing attribute value by the known value in another case that resembles as much as possible the case with the missing attribute value. In searching for the closest fit case we compare two vectors of attribute values, one vector corresponds to the case with a missing attribute value, the other vector is a candidate for the closest fit. The search is conducted for all cases, hence the name global closest fit. For each case a distance is computed, the case for which the distance is the smallest is the closest fitting case that is used to determine the missing attribute value. Let x and y be two cases. The distance between cases $x = (x_1, x_2, \cdots, x_n)$ and $y = (y_1, y_2, \cdots, y_n)$ is computed by $d(x, y) = \sum_{i=1}^{n} d(x_i, y_i)$, where $d(x_i, y_i) = 0$ for $x_i = y_i$ and $d(x_i, y_i) = |x_i - y_i|/r$ for $x_i \neq y_i$, where r is the difference between the maximum and minimum of the known values of the numerical attribute with a missing value.

(8) Concept Closest Fit: This method is similar to the global closest fit method. The difference is that the original data set, containing missing attribute values, is first split into smaller data sets, each smaller data set corresponds to a concept from the original data set. More precisely, every smaller data set is constructed from one of the original concepts, by restricting cases to the concept.

Base on Table 2, it is not difficult to obtain all possible values of the missing attribute value by Proposition 2.1, we omit these simple computation.

Proposition 2.2: All possible values of the missing attribute value in Table 2 are 1, 2 and 3.

III. Decision rule in information system

In order to use decision rule checking certainty of every value of the missing attribute value in Proposition 2.2, we need recall some basic concepts for information systems ([3], [4], [8]) and decision rules ([5], [6], [7]).

Definition 3.1: $S = (U, C \bigcup D, V, f)$ is called an information system.

(1) U, a nonempty finite set, is called the universe of discourse.

(2) $A = C \bigcup D$ is a finite set of attributes, where C and D are disjoint nonempty sets of condition attributes and decision attributes respectively.

(3) $f : U \times A \longrightarrow V$ is an information function.

(4) $V = \bigcup\{V_\alpha : \alpha \in A\}$, where $V_\alpha = \{f(u, \alpha) : u \in U\}$.

In particular, if $f(u, \alpha)$ is missing for some $u \in U$ and $\alpha \in A$, then $S = (U, C \bigcup D, V, f)$ is called an incomplete information system.

Remark 3.2: An information system $S = (U, C \bigcup D, V, f)$ can be denoted by a decision table. In this decision table, rows are labeled by elements of A, columns are labeled by elements of U, and $f(u, \alpha)$ lies in the cross of the column labeled by u and the row labeled by α.

Notation 3.3: (1) For a set B, $|B|$ denotes the cardinal of B.

(2) For a family of sets $\mathcal{F}_1, \mathcal{F}_2, \cdots, \mathcal{F}_k$, $\bigwedge\{\mathcal{F}_i : i = 1, 2, \cdots, k\} = \{\bigcap\{F_i : i = 1, 2, \cdots, k\} : F_i \in \mathcal{F}_i, i = 1, 2, \cdots, k\}$.

(3) Let R be an equivalence relation on a set U. U/R denotes the family consisting of all equivalence classes with respect to R and $[u]$ denotes the equivalence class with respect to R containing $u \in U$.

Definition 3.4: Let $S = (U, C \bigcup D, V, f)$ be an information system.

(1) For $a \in C \bigcup D$, we define an equivalence relation $\sim$ on U as follows:

$$u_i \sim u_j \Longleftrightarrow f(u_i, a) = f(u_j, a).$$

U/a denotes the family consisting of all equivalence classes with respect to $\sim$.

(2) For $B \subset C \bigcup D$, U/B denotes $\bigwedge\{U/b : b \in B\}$, which is a partition of U and induces an equivalence relation. For $x \in U$, the equivalence class of the partition U/B containing $x \in U$ is denoted by $B(x)$ and is called B-granule induced by x. In particular, $C(x)$ (resp. $D(x)$) are called the condition granule (resp. the decision granule) induced by x.

Definition 3.5: Let $S = (U, C \bigcup D, V, f)$ be an information system and $x \in U$. A sequence $f(x, c_1), f(x, c_2), \cdots, f(x, c_n), f(x, d_1), f(x, d_2), \cdots, f(x, d_m)$ is called a decision rule induced by x in $S = (U, C \bigcup D, V, f)$, where $\{c_1, c_2, \cdots, c_n\} = C$ and $\{d_1, d_2, \cdots, d_m\} = D$.

Remark 3.6: Let $S = (U, C \bigcup D, V, f)$ be an information system and $x \in U$. The decision rule induced by x in $S = (U, C \bigcup D, V, f)$ is denoted by $f(x, c_1), f(x, c_2), \cdots, f(x, c_n) \longrightarrow f(x, d_1), f(x, d_2), \cdots, f(x, d_m)$ or in short $C \longrightarrow_x D$.

Definition 3.7: Let $S = (U, C \bigcup D, V, f)$ be an information system and $x \in U$.

Put $\pi(C(x)) = \dfrac{|C(x)|}{|U|}$ and $\pi(D(x)) = \dfrac{|D(x)|}{|U|}$.

(1) The number $supp_x(C,D)$ is called the support of the decision rule $C \longrightarrow_x D$, where $supp_x(C,D) = |C(x) \bigcap D(x)|$.

(2) The number $\sigma_x(C,D)$ is called the strength of the decision rule $C \longrightarrow_x D$, where $\sigma_x(C,D) = \dfrac{supp_x(C,D)}{|U|}$.

(3) The number $cer_x(C,D)$ is called the certainty factor of the decision rule $C \longrightarrow_x D$, where $cer_x(C,D) = \dfrac{\sigma_x(C,D)}{\pi(C(x))}$.

(4) The number $cov_x(C,D)$ is called the coverage factor of the decision rule $C \longrightarrow_x D$, where $cov_x(C,D) = \dfrac{\sigma_x(C,D)}{\pi(D(x))}$.

Remark 3.8: Let $S = (U, C \bigcup D, V, f)$ be an information system and $x \in U$. Then the following hold from Definition 3.7.

(1) $cer_x(C,D) = \dfrac{supp_x(C,D)}{|C(x)|}$.

(2) $cov_x(C,D) = \dfrac{supp_x(C,D)}{|D(x)|}$.

Remark 3.9: Let $S = (U, C \bigcup D, V, f)$ be an information system and $x \in U$. Then $supp_x(C,D)$, $\sigma_x(C,D)$, $cer_x(C,D)$ and $cov_x(C,D)$ denote some degrees of condition attributes implying decision attributes for x.

IV. The incomplete information system $S = (U, C \bigcup D, V, f)$

This section establishes an incomplete information system corresponding to Table 2 by replacing all numbers in Table 2 by English alphabets. At first, we give some explanations for this incomplete information system.

Remark 4.1: Let $u_1, u_2, u_3, u_4, u_5, u_6, u_7$ denote 1978, 1980, 1990, 2000, 2006, 2007 and 2008, respectively. Put $U = \{u_1, u_2, u_3, u_4, u_5, u_6, u_7\}$. We take U as the universe of discourse.

Remark 4.2: Let $c_1, c_2, c_3, c_4, c_5, c_6, c_7, c_8$ denote cereals, meat, cotton lint, soybeans, groundnuts in shell, sugar cane, tea and fruit, respectively. Let d denotes gross domestic product. Put $C = \{c_1, c_2, c_3, c_4, c_5, c_6, c_7, c_8\}$ and $D = \{d\}$. We take C and D as the set of condition attributes and the set of a decision attribute, respectively.

Remark 4.3: the following classifications on attribute values in this incomplete information system are based on principle of statistics.

(1) c_{11} denotes that ranking of output of cereals in the world is 1; c_{12} denotes that ranking of output of cereals in the world is 2.

(2) c_{21} denotes that ranking of output of meat in the world is 1; c_{22} denotes that ranking of output of meat in the world is 3.

(3) c_{31} denotes that ranking of output of cotton lint in the world is 1; c_{32} denotes that ranking of output of cotton lint in the world is 2; c_{33} denotes that ranking of output of cotton lint in the world is 3.

(4) c_{41} denotes that ranking of output of soybeans in the world is 3; c_{42} denotes that ranking of output of soybeans in the world is 4.

(5) c_{51} denotes that ranking of output of groundnuts in shell in the world is 1; c_{52} denotes that ranking of output of groundnuts in shell in the world is 2.

(6) c_{61} denotes that ranking of output of sugar cane in the world is 3 or 4; c_{62} denotes that ranking of output of sugar cane in the world is 7; c_{63} denotes that ranking of output of sugar cane in the world is 9.

(7) c_{71} denotes that ranking of output of tea in the world is 1; c_{72} denotes that ranking of output of tea in the world is 2.

(8) c_{81} denotes that ranking of output of fruit in the world is 1; c_{82} denotes that ranking of output of fruit in the world is 4; c_{82} denotes that ranking of output of fruit in the world is 9 or 10.

(9) d_1 denotes that ranking of output of gross domestic product in the world is 3 or 4; d_2 denotes that ranking of output of gross domestic product in the world is 6; d_3 denotes that ranking of output of gross domestic product in the world is 10 or 11.

By Remark 4.1, Remark 4.2 and Remark 4.3, the incomplete information system $S = (U, C \bigcup D, V, f)$ corresponding to Table 2 is obtained and is denoted by the following incomplete decision table, where the missing attribute value is denoted by x.

TABLE III: INCOMPLETE DECISION TABLE II

U	u_1	u_2	u_3	u_4	u_5	u_6	u_7
c_1	c_{12}	c_{11}	c_{11}	c_{11}	c_{11}	c_{11}	c_{11}
c_2	c_{22}	c_{22}	c_{21}	c_{21}	c_{21}	c_{21}	c_{21}
c_3	c_{33}	c_{32}	c_{31}	c_{31}	c_{31}	c_{31}	x
c_4	c_{41}	c_{41}	c_{41}	c_{42}	c_{42}	c_{42}	c_{42}
c_5	c_{52}	c_{52}	c_{52}	c_{51}	c_{51}	c_{51}	c_{51}
c_6	c_{62}	c_{63}	c_{61}	c_{61}	c_{61}	c_{61}	c_{61}
c_7	c_{72}	c_{72}	c_{72}	c_{72}	c_{71}	c_{71}	c_{71}
c_8	c_{83}	c_{83}	c_{82}	c_{81}	c_{81}	c_{81}	c_{81}
d	d_3	d_3	d_3	d_2	d_1	d_1	d_1

Throughout the following sections, $S = (U, C \bigcup D, V, f)$ is always described by Table 3.

V. Check for values of the missing attribute value

By Proposition 2.2, all possible values of x in Table 3 are 1, 2 and 3. In this section, we will check the support, the strength, the certainty factor and the coverage factor of the decision rule for every value of x in Table 3.

Lemma 5.1: For the information system $S = (U, C \bigcup D, V, f)$, we have the following partitions of U.

(1) $U/c_1 = \{\{u_1\}, \{u_2, u_3, u_4, u_5, u_6, u_7\}\}$.

(2) $U/c_2 = \{\{u_1, u_2\}, \{u_3, u_4, u_5, u_6, u_7\}\}$.

(3) $U/c_3 = \{\{u_1\}, \{u_2\}, \{u_3, u_4, u_5, u_6, u_7\}\}$ for $x = c_{31}$;
$U/c_3 = \{\{u_1\}, \{u_2, u_7\}, \{u_3, u_4, u_5, u_6\}\}$ for $x = c_{32}$;
$U/c_3 = \{\{u_1, u_7\}, \{u_2\}, \{u_3, u_4, u_5, u_6\}\}$ for $x = c_{33}$.

(4) $U/c_4 = \{\{u_1, u_2, u_3\}, \{u_4, u_5, u_6, u_7\}\}$.

(5) $U/c_5 = \{\{u_1, u_2, u_3\}, \{u_4, u_5, u_6, u_7\}\}$.

(6) $U/c_6 = \{\{u_1\}, \{u_2\}, \{u_3, u_4, u_5, u_6, u_7\}\}$.

(7) $U/c_7 = \{\{u_1, u_2, u_3, u_4\}, \{u_5, u_6, u_7\}\}$.

(7) $U/c_8 = \{\{u_1, u_2\}, \{u_3\}, \{u_4, u_5, u_6, u_7\}\}$.

Based on Lemma 5.1, we have the following by some simple computation.

Lemma 5.2: For the information system $S = (U, C \bigcup D, V, f)$, we have the following partitions of U.

(1) $U/C = \{\{u_1\}, \{u_2\}, \{u_3\}, \{u_4\}, \{u_5, u_6, u_7\}\}$ for $x = c_{31}$.

$U/C = \{\{u_1\}, \{u_2\}, \{u_3\}, \{u_4\}, \{u_5, u_6\}, \{u_7\}\}$ for $x = c_{32}$ or $x = c_{33}$.

(2) $U/D = \{\{u_1, u_2, u_3\}, \{u_4\}, \{u_5, u_6, u_7\}\}$.

By Lemma 5.2, the following lemma can be obtained immediately.

Lemma 5.3: For the information system $S = (U, C \bigcup D, V, f)$, we have the following granules for u_7.

(1) $C(u_7) = \{u_5, u_6, u_7\}$ for $x = c_{31}$;

$C(u_7) = \{u_7\}$ for $x = c_{32}$ or $x = c_{33}$.

(2) $D(u_7) = \{u_5, u_6, u_7\}$.

Now we characterize the decision rule $C \longrightarrow_{u_7} D$ in $S = (U, C \bigcup D, V, f)$ by the support, the strength, the certainty factor and the coverage factor, which can be obtained from Definition 3.7, Remark 3.8 and Lemma 5.3.

Proposition 5.4: The following hold for the information system $S = (U, C \bigcup D, V, f)$

(1) $supp_{u_7}(C, D) = 3$ for $x = c_{31}$;

$supp_{u_7}(C, D) = 1$ for $x = c_{32}$ or $x = c_{33}$.

(2) $\sigma_{u_7}(C, D) = \dfrac{supp_{u_7}(C, D)}{|U|} = 3/7 \approx 0.429$ for $x = c_{31}$;

$\sigma_{u_7}(C, D) = \dfrac{supp_{u_7}(C, D)}{|U|} = 1/7 \approx 0.143$ for $x = c_{32}$ or $x = c_{33}$.

(3) $cer_{u_7}(C, D) = \dfrac{supp_{u_7}(C, D)}{|C(u_7)|} = 3/3 = 1.000$ for $x = c_{31}$;

$cer_{u_7}(C, D) = \dfrac{supp_{u_7}(C, D)}{|C(u_7)|} = 1/1 = 1.000$ for $x = c_{32}$ or $x = c_{33}$.

(4) $cov_{u_7}(C, D) = \dfrac{supp_{u_7}(C, D)}{|D(u_7)|} = 3/3 = 1.000$ for $x = c_{31}$;

$cov_{u_7}(C, D) = \dfrac{supp_{u_7}(C, D)}{|D(u_7)|} = 1/3 \approx 0.333$ for $x = c_{32}$ or $x = c_{33}$.

VI. Conclusion

For all values of the missing attribute value "x" in Proposition 2.2, the certainty factors of the decision rule $C \longrightarrow_{u_7} D$ are equivalent from Proposition 5.4(3). However, by Proposition 5.4(1),(2),(4) and Remark 3.9, the value of the missing attribute value "x" in Table 3 should be C_{31}. Thus, we obtain the following complete decision table, which improves Table 3 and answers Question 1.2.

TABLE IV: COMPLETE DECISION TABLE

U	u_1	u_2	u_3	u_4	u_5	u_6	u_7
c_1	c_{12}	c_{11}	c_{11}	c_{11}	c_{11}	c_{11}	c_{11}
c_2	c_{22}	c_{22}	c_{21}	c_{21}	c_{21}	c_{21}	c_{21}
c_3	c_{33}	c_{32}	c_{31}	c_{31}	c_{31}	c_{31}	c_{31}
c_4	c_{41}	c_{41}	c_{41}	c_{42}	c_{42}	c_{42}	c_{42}
c_5	c_{52}	c_{52}	c_{52}	c_{51}	c_{51}	c_{51}	c_{51}
c_6	c_{62}	c_{63}	c_{61}	c_{61}	c_{61}	c_{61}	c_{61}
c_7	c_{72}	c_{72}	c_{72}	c_{72}	c_{71}	c_{71}	c_{71}
c_8	c_{83}	c_{83}	c_{82}	c_{81}	c_{81}	c_{81}	c_{81}
d	d_3	d_3	d_3	d_2	d_1	d_1	d_1

By Remark 4.3(3), c_{31} denotes that ranking of output of cotton lint in the world is 1, so the missing value in Table 1 should be 1. In the end of this paper, we give the following complete data table by filling the gap in Table 1, which answers Question 1.1 and gives a further improvement for United Nations FAO Database.

TABLE V: COMPLETE DATA TABLE

Item	u_1	u_2	u_3	u_4	u_5	u_6	u_7
Cereals	2	1	1	1	1	1	1
Meat	3	3	1	1	1	1	1
Cotton Lint	3	2	1	1	1	1	1
Soybeans	3	3	3	4	4	4	4
Groundnuts in Shell	2	2	2	1	1	1	1
Sugar Cane	7	9	4	3	3	3	3
Tea	2	2	2	1	1	1	1
Fruit	9	10	4	1	1	1	4

ACKNOWLEDGMENT

This work is supported by the National Natural Science Foundation of China (No. 11226085 and 11061004).

REFERENCES

[1] National Bureau of Statistics of China, *ranking lists of Chinese output of the main agricultural commodity in the world*, http://www.stats.gov.cn/ 2011.
[2] National Bureau of Statistics of China, *ranking list of Chinese the main index in the world*, http://www.stats.gov.cn/ 2011.
[3] J. W. Grzymala-Busse, Rough Sets and Granular Computing in Dealing with Missing Attribute Values, in: Handbook of Granular Computing, Edited by W. Pedrycz, A. Skowron and V. Kreinovich, John Wiley & Sons, Ltd., 2008, 873-888.
[4] Z. Pawlak, Rough sets, International Journal of Computer and Information Sciences, 11(1982), 341-356.
[5] Z. Pawlak, Decisions rules and flow networks, European Journal of Operational Research, 154(2004), 184-190.
[6] Z. Pawlak, Can Bayesian confirmation measures be useful for rough set decision rules? Engineering Applications of Artificial Intelligence, 17(2004), 345-361.
[7] Z. Pawlak, Some remarks on conflict analysis, European Journal of Operational Research, 166(2005), 649-654.
[8] A. Skowron, J. F. Peters, Rough-Granular Computing, in: Handbook of Granular Computing, Edited by W. Pedrycz, A. Skowron and V. Kreinovich, John Wiley & Sons, Ltd., 2008, 285-328.
[9] United Nations Development Porgram, *Human Development Report 2009/2010*, United Nations UNSD Database 2011.
[10] Food and Agriculture Organization of the United Nations, Agricultural Commodity Statistics Yearbook, United Nations FAO Database, 2011.