Data Filling Approach of Soft Sets under Incomplete Information

Hongwu Qin, Xiuqin Ma, Tutut Herawan, and Jasni Mohamad Zain

Faculty of Computer Systems and Software Engineering
Universiti Malaysia Pahang
Lebuh Raya Tun Razak, Gambang 26300, Kuantan, Malaysia
qhwump@gmail.com, xueener@yahoo.com.cn,
tutut@ump.edu.my, jasni@ump.edu.my

Abstract. Incomplete information in a soft set restricts the usage of the soft set. To make the incomplete soft set more useful, in this paper, we propose a data filling approach for incomplete soft set in which missing data is filled in terms of the association degree between the parameters when stronger association exists between the parameters or in terms of the probability of objects appearing in the mapping sets of parameters when no stronger association exists between the parameters. An illustrative example is employed to show the feasibility and validity of our approach in practical applications.

Keywords: Soft sets, Incomplete soft sets, Data filling, Association degree.

1 Introduction

In 1999, Molodtsov [1] proposed soft set theory as a new mathematical tool for dealing with vagueness and uncertainties. At present, work on the soft set theory is progressing rapidly and many important theoretical models have been presented, such as soft groups [2], soft rings [3], soft semirings [4], soft ordered semigroup [5] and exclusive disjunctive soft sets [6]. The research on fuzzy soft set has also received much attention since its introduction by Maji et al. [7]. Several extension models including intuitionistic fuzzy soft sets [8], interval-valued fuzzy soft sets [9] and interval-valued intuitionistic fuzzy soft set [10] are proposed in succession. At the same time, researchers have also successfully applied soft sets to deal with some practical problems, such as decision making [11-14], economy forecasting [15], maximal association rules mining [16], etc.

The soft sets mentioned above, either in theoretical study or practical applications are based on complete information. However, incomplete information widely exists in practical problems. For example, an applicant perhaps misses age when he/she fills out an application form. Missing or unclear data often appear in questionnaire due to the fact that attendees give up some questions or can not understand the meaning of questions well. In addition, other reasons like mistakes in the process of measuring and collecting data, restriction of data collecting also can cause unknown or missing data. Hence, soft sets under incomplete information become incomplete soft sets. In order to handle incomplete soft sets, new data processing methods are required.

N.T. Nguyen, C.-G. Kim, and A. Janiak (Eds.): ACIIDS 2011, LNAI 6592, pp. 302–311, 2011.
© Springer-Verlag Berlin Heidelberg 2011

Yan and Zhi [17] initiated the study on soft sets under incomplete information. They put forward improved data analysis approaches for standard soft sets and fuzzy soft sets under incomplete information, respectively. For crisp soft sets, the decision value of an object with incomplete information is calculated by weighted-average of all possible choice values of the object, and the weight of each possible choice value is decided by the distribution of other available objects. Incomplete data in fuzzy soft sets is predicted based on the method of average probability. However, there is inherent deficiency in their method. For crisp soft sets, directly calculating the decision value of an object with incomplete information makes the method only applicable to decision making problems. During the process of data analysis the soft sets keep invariable, in other words the missing data is still missing. Therefore, the soft sets can not be used in other fields but decision making.

Intuitively, there are two methods which can be used to overcome the deficiency in [17]. The simplest method is deletion that the objects with incomplete data will be deleted directly from incomplete soft sets. This method, however, probably makes valuable information missing. Another method is data filling, that is, the incomplete data will be estimated or predicted based on the known data. Data filling converts an incomplete soft set into a complete soft set, which makes the soft set more useful. So far, few researches focus on data filling approaches for incomplete soft sets.

In this paper, we propose a data filling approach for incomplete soft sets. We analyze the relations between the parameters and define the notion of association degree to measure the relations. In our method, we give priority to the relations between the parameters due to its higher reliability. When the mapping set of a parameter includes incomplete data, we firstly look for another parameter which has the stronger association with the parameter. If another parameter is found, the missing data in the mapping set of the parameter will be filled according to the value in the corresponding mapping set of another parameter. If no parameter has the stronger association with the parameter, the missing data will be filled in terms of the probability of objects appearing in the mapping set of the parameter. There are two main contributions in this work. First, we present the applicability of the data filling method to handle incomplete soft sets. Second, we introduce the relation between parameters to fill the missing data.

The rest of this paper is organized as follows. The following section presents the notions of soft sets and incomplete soft sets. Section 3 analyzes the relation between the parameters of soft set and defines the notion of association degree to measure the relation. In Section 4, we present our algorithm for filling the missing data and give an illustrative example. Finally, conclusions are given in Section 5.

2 Preliminaries

Let U be an initial universe of objects, E be the set of parameters in relation to objects in U , $P(U)$ denote the power set of U . The definition of soft set is given as follows.

Definition 2.1 ([1]). A pair (F,E) is called *a soft set* over U , where F is a mapping given by

$$F: E \rightarrow P(U)$$

From definition, a soft set (F, E) over the universe U is a parameterized family of subsets of the universe U, which gives an approximate description of the objects in U. For any parameter $e \in E$, the subset $F(e) \subseteq U$ may be considered as the set of e-approximate elements in the soft set (F, E).

Example 1. Let us consider a soft set (F, E) which describes the "attractiveness of houses" that Mr. X is considering to purchase. Suppose that there are six houses in the univers $U = \{h_1, h_2, h_3, h_4, h_5, h_6\}$ under consideration and $E = \{e_1, e_2, e_3, e_4, e_5\}$ is the parameter set, where $e_i (i = 1,2,3,4,5)$ stands for the parameters "beautiful", "expensive", "cheap", "good location" and "wooden" respectively. Consider the mapping $F: E \rightarrow P(U)$ given by "houses (.)", where (.) is to be filled in by one of parameters $e \in E$. Suppose that $F(e_1) = \{h_1, h_3, h_6\}$, $F(e_2) = \{h_1, h_2, h_3, h_6\}$, $F(e_3) = \{h_4, h_5\}$, $F(e_4) = \{h_1, h_2, h_6\}$, $F(e_5) = \{h_5\}$. Therefore, $F(e_1)$ means "houses (beautiful)", whose value is the set $\{h_1, h_3, h_6\}$.

In order to facilitate storing and dealing with soft set, the binary tabular representation of soft set is often given in which the rows are labeled by the object names and columns are labeled by the parameter names, and the entries are $F(e_j)(x_i), (e_j \in E, x_i \in U, j = 1,2,..m, x = 1,2,..n)$. If $x_i \in F(e_j)$, then $F(e_j)(x_i) = 1$, otherwise $F(e_j)(x_i) = 0$. Table 1 is the tabular representation of the soft set (F, E) in Example 1.

Table 1. Tabular representation of the soft set (F, E)

U	e_1	e_2	e_3	e_4	e_5
h_1	1	1	0	1	0
h_2	0	1	0	1	0
h_3	1	1	0	0	0
h_4	0	0	1	0	0
h_5	0	0	1	0	1
h_6	1	1	0	1	0

Definition 2.2. A pair (F, E) is called *an incomplete soft set* over U, if there exists $x_i \in U(i = 1,2...,n)$ and $e_j \in E(j = 1,2...,m)$, making $x_i \in F(e_j)$ unknown, that is, $F(e_j)(x_i) = null$.

In tabular representation, null is represented by "*".

Example 2. Assume a community college is recruiting some new teachers and there are 8 persons applied for the job. Let us consider a soft set (F,E) which describes the "capability of the candidates". The universe $U=\{c_1,c_2,c_3,c_4,c_5,c_6,c_7,c_8\}$ and $E=\{e_1,e_2,e_3,e_4,e_5,e_6\}$ is the parameter set, where $e_i(i=1,2,3,4,5,6)$ stands for the parameters "experienced", "young age", "married", "the highest academic degree is Doctor", "the highest academic degree is Master" and "studied abroad" respectively. Consider the mapping $F:E\to P(U)$ given by "candidates (.)", where (.) is to be filled in by one of parameters $e\in E$. Suppose that

$$F(e_1)=\{c_1,c_2,c_5,c_7\},\ F(e_2)=\{c_3,c_4,c_6\},\ F(e_3)=\{c_1,c_5,c_7,c_8\},$$
$$F(e_4)=\{c_2,c_4,c_5,c_8\},\quad F(e_5)=\{c_1,c_3,c_6,c_7\},\ F(e_6)=\{c_8\}.$$

Therefore, $F(e_1)$ means "candidates (experienced)", whose value is the set $\{c_1,c_2,c_5,c_7\}$. Unfortunately, several applicants missed some information. As a result, the soft set (F,E) becomes an incomplete soft set. Table 2 is the tabular representation of the incomplete soft set (F,E). If $c_j\in F(e_i)$ is unknown, $F(e_i)(c_j)='*'$, where $F(e_i)(c_j)$ are the entries in Table 2.

Table 2. Tabular representation of the incomplete soft set (F,E)

U	e_1	e_2	e_3	e_4	e_5	e_6
c_1	1	0	1	0	1	0
c_2	1	0	0	1	0	0
c_3	0	1	0	0	1	0
c_4	0	1	*	1	0	*
c_5	1	0	1	1	0	0
c_6	0	1	0	0	*	0
c_7	1	*	1	0	1	0
c_8	0	0	1	1	0	0

3 Association Degree between Parameters in an Incomplete Soft Set

So far, few research focus on the associations between parameters in the soft sets. Actually, for one object, there always exist some obvious or hidden associations between parameters. This is just like for a person, as we know, the attribute weight has some certain relation with the attribute height.

Let us reconsider Example 1 and Example 2. There are many obvious associations in the two examples. In Example 1, it is easy to find that if a house is expensive, the house is not cheap, vice versa. There is inconsistent association between parameter "expensive" and parameter "cheap". Generally speaking, if a house is beautiful or has a good location, the house is expensive. There is consistent association between parameter "beautiful" and parameter "expensive" or between parameter "good location" and parameter "expensive". Similarly, in Example 2, there is obvious inconsistent association between parameter "the highest academic degree is Doctor" and parameter "the highest academic degree is Master". A candidate has only one highest academic degree. We can also find that if a candidate is experienced or has been married, in general, he/she is not young. There is inconsistent association between parameter "experienced" and parameter "young age" or between parameter "married" and parameter "young age".

These associations reveal the interior relations of an object. In a soft set, these associations between parameters will be very useful for filling incomplete data. If we have already found that parameter e_i is associated with parameter e_j and there are missing data in $F(e_i)$, we can filling the missing data according to the corresponding data in $F(e_j)$ based on the association between e_i and e_j. To measure these associations, we define the notion of association degree and some relative notions.

Let U be a universe set and E be a set of parameters. U_{ij} denotes the set of objects that have specified values 0 or 1 both on parameter e_i and parameter e_j such that

$$U_{ij} = \{x \mid F(e_i)(x) \neq '*' \quad and \quad F(e_j)(x) \neq '*', x \in U\}$$

In other words, U_{ij} stands for the set of objects that have known data both on e_i and e_j. Based on U_{ij}, we have the following definitions.

Definition 3.1. Let E be a set of parameters and $e_i, e_j \in E$, $(i, j = 1,2,...m)$. *Consistent Association Number* between parameter e_i and parameter e_j is denoted by CN_{ij} and defined as

$$CN_{ij} = \left|\{x \mid F(e_i)(x) = F(e_j)(x), x \in U_{ij}\}\right|$$

where m denotes the number of parameters, $|.|$ denotes the cardinality of set.

Definition 3.2. Let E be a set of parameters and $e_i, e_j \in E$, $(i, j = 1,2,...m)$. *Consistent Association Degree* between parameter e_i and parameter e_j is denoted by CD_{ij} and defined as

$$CD_{ij} = \frac{CN_{ij}}{|U_{ij}|}$$

Obviously, the value of CD_{ij} is in [0, 1]. Consistent Association Degree measures the extent to which the value of parameter e_i keeps consistent with that of parameter e_j over U_{ij}.

Similarly, we can define *Inconsistent Association Number* and *Inconsistent Association Degree* as follows.

Definition 3.3. Let E be a set of parameters and $e_i, e_j \in E$, $(i, j = 1,2,...m)$. *Inconsistent Association Number* between parameter e_i and parameter e_j is denoted by IN_{ij} and defined as

$$IN_{ij} = \left|\left\{x \mid F(e_i)(x) \neq F(e_j)(x), x \in U_{ij}\right\}\right|$$

Definition 3.4. Let E be a set of parameters and $e_i, e_j \in E$, $(i, j = 1,2,...m)$. *Inconsistent Association Degree* between parameter e_i and parameter e_j is denoted by ID_{ij} and defined as

$$ID_{ij} = \frac{IN_{ij}}{\left|U_{ij}\right|}$$

Obviously, the value of ID_{ij} is also in [0, 1]. Inconsistent Association Degree measures the extent to which parameters e_i and e_j is inconsistent.

Definition 3.5. Let E be a set of parameters and $e_i, e_j \in E$, $(i, j = 1,2,...m)$. *Association Degree* between parameter e_i and parameter e_j is denoted by D_{ij} and defined as

$$D_{ij} = \max\left\{CD_{ij}, ID_{ij}\right\}$$

If $CD_{ij} > ID_{ij}$, then $D_{ij} = CD_{ij}$, it means most of objects over U_{ij} have consistent values on parameters e_i and e_j. If $CD_{ij} < ID_{ij}$, then $D_{ij} = ID_{ij}$, it means most of objects over U_{ij} have inconsistent values on parameters e_i and e_j. If $CD_{ij} = ID_{ij}$, it means that there is the lowest association degree between parameters e_i and e_j.

Property 3.1. For any parameters and e_j, $D_{ij} \geq 0.5$. $(i, j = 1,2,...m)$.

Proof. For any parameters e_i and e_j, from the definitions of CD_{ij} and ID_{ij}, we have

$$CD_{ij} + ID_{ij} = 1.$$

Therefore, at least one of CD_{ij} and ID_{ij} is more than 0.5, namely, $D_{ij} = \max\left\{CD_{ij}, ID_{ij}\right\} \geq 0.5$. □

Definition 3.6. Let E be a set of parameters and $e_i \in E$ $(i = 1,2,...m)$. *Maximal Association Degree* of parameter e_i is denoted by D_i and defined as

$$D_i = \max D_{ij}, \quad \mathrm{j} = 1,2,...\mathrm{m}.$$

where m is the number of parameters.

4 The Algorithm for Data Filling

In terms of the analysis in the above section, we can propose the data filling method based on the association degree between the parameters. Suppose the mapping set $F(e_i)$ of parameter e_i includes missing data. At first, calculate association degrees between parameter e_i and each of other parameters respectively over existing complete information, and then find the parameter e_j which has the maximal association degree with parameter e_i. Finally the missing data in $F(e_i)$ will be filled according to the corresponding data in mapping set $F(e_j)$. However, sometimes a parameter perhaps has a lower maximal association degree, that is, the parameter has weaker association with other parameters. In this case, the association is not reliable any more and we have to find other methods. Inspired by the data analysis approach in [17], we can use the probability of objects appearing in the $F(e_i)$ to fill the missing data. In our method we give priority to the association between the parameters instead of the probability of objects appearing in the $F(e_i)$ to fill the missing data due to the fact that the relation between the parameters are more reliable than that between the objects in soft set. Therefore, we can set a threshold, if the maximal association degree equals or exceeds the predefined threshold, the missing data in $F(e_i)$ will be filled according to the corresponding data in $F(e_j)$, or else the missing data will be filled in terms of the probability of objects appearing in the $F(e_i)$. Fig. 1 shows the details of the algorithm.

In order to make the computation of association degree easier, we construct an association degree table in which rows are labeled by the parameters including missing data and columns are labeled by all of the parameters in parameter set, and the entries are association degree D_{ij}. To distinguish the inconsistent association degree from consistent degree, we add a minus sign before the inconsistent association degree.

Example 3. Reconsider the incomplete soft set (F, E) in Example 2. There are missing data in $F(e_2)$, $F(e_3)$, $F(e_5)$ and $F(e_6)$. We will fill the missing data in (F, E) by using Algorithm 1. Firstly, we construct an association degree table as Table 3.

For parameter e_2, we can see from the table, the association degree $D_{21} = 0.86$, $D_{23} = 0.83$, $D_{24} = 0.71$, $D_{25} = 0.67$, $D_{26} = 0.67$, where D_{21}, D_{23} and D_{24} are

```
Algorithm 1.
1. Input the incomplete soft set $(F,E)$.
2. Find $e_i$, which includes missing data $F(e_i)(x)$.
3. Compute $D_{ij}$, $j=1,2,...,m$, where $m$ is the number of parameters in $E$.
4. Compute maximal association degree $D_i$.
5. If $D_i \geq \lambda$, find the parameter $e_j$ which has the maximal association degree $D_i$ with
   parameter $e_i$.
6. If there is consistent association between $e_i$ and $e_j$, $F(e_i)(x)=F(e_j)(x)$. If there is
   inconsistent association between $e_i$ and $e_j$, $F(e_i)(x)=1-F(e_j)(x)$.
7. If $D_i < \lambda$, compute the probabilities $P_1$ and $P_0$ that stand for object $x$ belongs to and does not
   belong to $F(e_i)$, respectively.
       $P_1 = \frac{n_1}{n_1+n_0}$,  $P_0 = \frac{n_0}{n_1+n_0}$
   where $n_1$ and $n_0$ stand for the number of objects that belong to and does not belong to $F(e_i)$,
   respectively.
8. If $P_1 > P_0$, $F(e_i)(x)=1$. If $P_0 > P_1$, $F(e_i)(x)=0$. If $P_1 = P_0$, 0 or 1 may be assigned
   to $F(e_i)(x)$.
9. If all of the missing data is filled, algorithm end, or else go to step 2.
```

Fig. 1. The algorithm for data filling

from inconsistent association degree, D_{25} and D_{26} are from consistent association degree. The maximal association degree $D_2 = 0.86$. We set the threshold $\lambda = 0.8$. Therefore, in terms of the Algorithm 1, we can fill $F(e_2)(c_7)$ according to $F(e_1)(c_7)$. Because $F(e_1)(c_7) = 1$ and there is inconsistent association between parameters e_2 and e_1, so we fill 0 into $F(e_2)(c_7)$. Similarly, we can fill 0, 1 into $F(e_3)(c_4)$ and $F(e_5)(c_6)$ respectively.

Table 3. Association degree table for incomplete soft set (F,E)

	e_1	e_2	e_3	e_4	e_5	e_6
e_2	-0.86	-	-0.83	-0.71	0.67	0.67
e_3	0.71	-0.83	-	0.57	0.5	-0.57
e_5	0.57	0.67	0.5	-1	-	0.5
e_6	-0.57	0.67	0.57	0.57	0.5	-

For parameter e_6, we have the maximal association degree $D_6 = 0.67 < \lambda$. That means there is not reliable association between parameter e_6 and other parameters. So we can not fill the data $F(e_6)(c_4)$ according to other parameters. In terms of the steps 8 and 9 in Algorithm 1, we have $P_0 = 1$, $P_1 = 0$. Therefore, we fill 0 into $F(e_6)(c_4)$. Table 4 shows the tabular representation of the filled soft set (F, E).

Table 4. Tabular representation of the incomplete soft set (F, E)

U	e_1	e_2	e_3	e_4	e_5	e_6
c_1	1	0	1	0	1	0
c_2	1	0	0	1	0	0
c_3	0	1	0	0	1	0
c_4	0	1	0	1	0	0
c_5	1	0	1	1	0	0
c_6	0	1	0	0	1	0
c_7	1	0	1	0	1	0
c_8	0	0	1	1	0	0

5 Conclusion

In this paper, we propose a data filling approach for incomplete soft sets. We analyze the relations between the parameters and define the notion of association degree to measure the relations. If the mapping set of a parameter includes incomplete data, we firstly look for another parameter which has the stronger association with the parameter. If another parameter is found, the missing data in the mapping set of the parameter will be filled according to the value in the corresponding mapping set of another parameter. If no parameter has the stronger association with the parameter, the missing data will be filled in terms of the probability of objects appearing in the mapping set of the parameter. We validate the method by an example and draw conclusion that data filling method is applicable to handle incomplete soft sets and the relations between parameters can be applied to fill the missing data. The method can be used to handle various applications involved incomplete soft sets.

Acknowledgments. This work was supported by PRGS under the Grant No. GRS100323, Universiti Malaysia Pahang, Malaysia.

References

1. Molodtsov, D.: Soft set theory_First results. Computers and Mathematics with Applications 37, 19–31 (1999)
2. Aktas, H., Cagman, N.: Soft sets and soft groups. Information Sciences 177, 2726–2735 (2007)

3. Acar, U., Koyuncu, F., Tanay, B.: Soft sets and soft rings. Computers and Mathematics with Applications 59, 3458–3463 (2010)
4. Feng, F., Jun, Y.B., Zhao, X.: Soft semirings. Computers and Mathematics with Applications 56, 2621–2628 (2008)
5. Jun, Y.B., Lee, K.J., Khan, A.: Soft ordered semigroups. Math. Logic Quart. 56, 42–50 (2010)
6. Xiao, Z., Gong, K., Xia, S., Zou, Y.: Exclusive disjunctive soft sets. Computers and Mathematics with Applications 59, 2128–2137 (2010)
7. Maji, P.K., Biswas, R., Roy, A.R.: Fuzzy soft sets. Journal of Fuzzy Mathematics 9, 589–602 (2001)
8. Maji, P.K., Biswas, R., Roy, A.R.: Intuitionistic fuzzy soft sets. Journal of Fuzzy Mathematics 9, 677–692 (2001)
9. Yang, X.B., Lin, T.Y., Yang, J., Dongjun, Y.L.A.: Combination of interval-valued fuzzy set and soft set. Computers and Mathematics with Applications 58, 521–527 (2009)
10. Jiang, Y., Tang, Y., Chen, Q., Liu, H., Tang, J.: Interval-valued intuitionistic fuzzy soft sets and their properties. Computers and Mathematics with Applications 60, 906–918 (2010)
11. Maji, P.K., Roy, A.R.: An application of soft sets in a decision making problem. Computers and Mathematics with Applications 44, 1077–1083 (2002)
12. Feng, F., Jun, Y.B., Liu, X., Li, L.: An adjustable approach to fuzzy soft set based decision making. Journal of Computational and Applied Mathematics 234, 10–20 (2010)
13. Feng, F., Li, Y., Leoreanu-Fotea, V.: Application of level soft sets in decision making based on interval-valued fuzzy soft sets. Computers and Mathematics with Applications 60, 1756–1767 (2010)
14. Jiang, Y., Tang, Y., Chen, Q.: An adjustable approach to intuitionistic fuzzy soft sets based decision making. Applied Mathematical Modelling 35, 824–836 (2011)
15. Xiao, Z., Gong, K., Zou, Y.: A combined forecasting approach based on fuzzy soft sets. Journal of Computational and Applied Mathematics 228, 326–333 (2009)
16. Herawan, T., Mat Deris, M.: A soft set approach for association rules mining. Knowledge-Based Systems 24, 186–195 (2011)
17. Zou, Y., Xiao, Z.: Data analysis approaches of soft sets under incomplete information. Knowledge-Based Systems 21, 941–945 (2008)