Int. J. Appl. Math. Comput. Sci., 2012, Vol. 22, No. 4, 817–828
DOI: 10.2478/v10006-012-0060-3

DFIS: A NOVEL DATA FILLING APPROACH FOR AN INCOMPLETE SOFT SET

HONGWU QIN [*], XIUQIN MA [**,*], TUTUT HERAWAN [**], JASNI MOHAMAD ZAIN [**]

[*] College of Computer Science and Engineering
Northwest Normal University, Lanzhou Gansu, 730070, China
e-mail: {qhwump,xueener}@gmail.com

[**] Faculty of Computer Systems and Software Engineering
University of Malaysia, Pahang, Lebuh Raya Tun Razak, Gambang 26300, Kuantan, Malaysia
e-mail: {tutut,jasni}@ump.edu.my

The research on incomplete soft sets is an integral part of the research on soft sets and has been initiated recently. However, the existing approach for dealing with incomplete soft sets is only applicable to decision making and has low forecasting accuracy. In order to solve these problems, in this paper we propose a novel data filling approach for incomplete soft sets. The missing data are filled in terms of the association degree between the parameters when a stronger association exists between the parameters or in terms of the distribution of other available objects when no stronger association exists between the parameters. Data filling converts an incomplete soft set into a complete soft set, which makes the soft set applicable not only to decision making but also to other areas. The comparison results elaborated between the two approaches through UCI benchmark datasets illustrate that our approach outperforms the existing one with respect to the forecasting accuracy.

Keywords: soft sets, incomplete soft sets, data filling, association degree.

1. Introduction

A lot of practical and complicated problems in many fields involve uncertain, fuzzy, not clearly defined data. A wide variety of theories are applicable to modeling vagueness as diverse as probability theory, fuzzy sets (Zadeh, 1965; Li and Chiang, 2011), rough sets (Pawlak, 1982; Zhong and Skowron, 2001), intuitionistic fuzzy sets (Atanassov, 1986), vague sets (Gau and Buehrer, 1993) and interval mathematics (Gorzalzany, 1987), each of which has its inherent difficulties as pointed out by Molodtsov (2004). To overcome these difficulties, Molodtsov (1999) proposed soft set theory as a new mathematical tool for dealing with vagueness and uncertainties. A soft set is a parameterized family of the subsets of a universal set. It can be said that soft sets are neighborhood systems and a special case of context-dependent fuzzy sets. In contrast to all these theories, soft set theory is free from the above limitations and has no problem of setting the membership function, which makes it very convenient and easy to apply in practice. Therefore, it has a rich potential for applications in several directions, some of which had already been demonstrated by Molodtsov (1999), such as the smoothness of functions, game theory, operations research, Riemann integration, Perron integration, probability theory, and measure theory.

Presently, research on the soft set theory is progressing rapidly. Maji *et al*. (2003) firstly introduced some definitions of the related operations on soft sets. Ali *et al*. (2009) took into account some errors of former studies and put forward some new operations on soft sets. Maji and Roy (2002) employed soft sets to solve the decision-making problem. Chen *et al*. (2005) pointed out that the conclusion of soft set reduction offered by Maji and Roy (2002) was incorrect, and then presented a new notion of parameterization reduction in soft sets in comparison with the definition to the related concept of attribute reduction in rough set theory. The concept of normal parameter reduction is introduced by Kong *et al*. (2008), who overcame the problem of suboptimal choice and added a parameter set of soft sets. An algorithm for normal parameter reduction was also presented. However, the algorithm is hard to understand and involves a great amount of computation. To improve this algorithm, Ma *et al*. (2011) proposed a new efficient normal parameter reduction algorithm of soft sets. Soft set theory can also be applied to data mining. An alternative approach to

mining regular association rules and maximal association rules from a transactional dataset using soft set theory was presented by Herawan and Mat Deris (2011).

Furthermore, the soft set model can also be combined with other mathematical models. Therefore the definitions of soft groups (Aktas and Cagman, 2007), soft ideals and idealistic soft BCK/BCI-algebras (Jun and Park, 2008), soft semirings, soft subsemirings, soft ideals and idealistic soft semirings (Feng *et al*., 2008) have been given. Çağman and Enginoğlu (2010) defined soft matrices and their operations and described products of soft matrices and their properties. Qin and Kong (2010) introduced the concept of soft equality and derived some related properties. Several extension models including vague soft sets (Xu *et al*., 2010), fuzzy soft sets (Maji *et al*., 2001a; Majumdar and Samanta, 2010), intuitionistic fuzzy soft sets (Maji, 2009; Maji *et al*., 2001b; 2004), interval-valued fuzzy soft sets (Yang *et al*., 2009) and interval-valued intuitionistic fuzzy soft sets (Jiang *et al*., 2010) are proposed in succession. It could be shown that soft set theory is closely associated with rough sets (Pei and Miao, 2005; Herawan and Mat Deris, 2009; Feng *et al*., 2009; Feng, 2009). Based on these extension models, some applications to decision making (Maji and Roy, 2007; Kong *et al.*, 2009; Feng, 2010a; 2010b; Jiang *et al.*, 2011; Qin *et al.*, 2011a; 2011b) and the combined forecasting approach (Xiao *et al*., 2009) were shown.

The soft sets mentioned above, either in theoretical studies 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 (Nowicki, 2010) often appear in questionnaires due to the fact that attendees give up some questions or may not understand the meaning of questions correctly. In addition, other reasons like mistakes in the process of measuring and collecting data or restriction of data collecting can also result in unknown or missing data. Hence, soft sets under incomplete information become incomplete soft sets.

There are some traditional approaches for dealing with incomplete information. The simplest approach is to directly delete unknown or missing data from incomplete information systems, which will, however, lead to the missing of some valuable information. Data filling is another approach for dealing with incomplete information, which can predict inexistent or missing data by evidence theory, expert experience, average and Bayesian models, and so on. However, it is necessary for evidence theory and Bayesian models to learn the evidence function and the probability distribution in advance. Also, experts' experience is not objective. Data analysis approaches to rough sets in an incomplete information system are also described by Thiesson (1995), Zhang and Li (2006), as well as Quinlan (1989). It is known that the value domain of a soft set is a set of subsets of all objects in an initial finite universe. This particularity of value domains of mapping functions in soft sets makes all of these traditional methods mentioned above inapplicable directly applied to dealing with incomplete soft sets. In order to handle incomplete soft sets, new data processing methods are required.

Zou and Xiao (2008) 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. For crisp soft sets, the decision value of an object with incomplete information is calculated by the 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 are two inherent deficiencies in their method. Firstly, 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, soft sets remain invariable; in other words, the missing data are still missing. Therefore, soft sets cannot be used in other fields but decision making. Secondly, in the decision making problem, all of objects are competitive and each choice value of objects is independent of that of other objects. Thus the distribution of other available objects deciding the weight of each possible choice value is inappropriate to deal with incomplete soft sets, which makes this method of a low accuracy.

In order to overcome the two inherent deficiencies of Zou and Xiao (2008), in this paper we propose a novel data filling approach for incomplete soft sets (called DFIS). We firstly define the notion of the association degree to measure the relations between the parameters, which is more reliable than the distribution of other available objects. The missing data are filled in terms of the association degree between the parameters when a stronger association exists between the parameters or in terms of the probability of other available objects when no stronger association exists between the parameters. There are two main contributions in this work. First, we present the applicability of the data filling method to handle incomplete soft sets. Data filling converts an incomplete soft set into a complete soft set, which makes the soft set more useful. Second, we introduce the association degree between parameters to fill the missing data, which can improve the accuracy compared with the method of Zou and Xiao (2008).

The remainder of this paper is organized as follows. The following section presents the notions of soft sets and incomplete soft sets. Section 3 studies the data analysis approaches of soft sets under incomplete information

put forward by Zou and Xiao (2008). Section 4 defines the notion of the association degree between the parameters and presents a novel data filling approach for incomplete soft sets. In Section 5, a comparison between two approaches is elaborated through a Boolean data set, and then the experimental results are analyzed and comparisons are done based on five UCI benchmark datasets. Finally, conclusions are given in Section 6.

2. Preliminaries

In this section, we review some definitions and properties regarding soft sets. Let U be a non-empty initial universe of objects, E be a set of parameters in relation to objects in U, $P(U)$ be the power set of U, and $A \subset E$. The definition of a soft set is given as follows.

Definition 1. (*Molodtsov, 2004*) A pair (F, A) is called a *soft set* over U, where F is a mapping given by $F : A \to P(U)$. That is, a soft set over U is a parameterized family of subsets of the universe U.

Definition 2. An *information system* is a quadruple $S = (U, A, V, f)$, where $U = \{u_1, u_2, \ldots, u_{|U|}\}$ is a non-empty finite set of objects, $A = \{a_1, a_2, \ldots, a_{|A|}\}$ is a non-empty finite set of attributes, $V = \bigcup_{a \in A} V_a$, V_a is the domain (value set) of the attribute a, $f : U \times A \to V$ is an information function such that $f(u, a) \in V_a$, for every $(u, a) \in U \times A$, called an information (knowledge) function.

An information system is also called a knowledge representation system or an attribute-valued system and can be intuitively expressed in terms of an information table. In an information system $S = (U, A, V, f)$, if $V_a = \{0, 1\}$ for every $a \in A$, then S is called a *Boolean-valued information system*.

Proposition 1. (Thiesson, 1995) *If (F, E) is a soft set over the universe U, then (F, E) is a Boolean valued information system $S = (U, A, V_{\{0,1\}}, f)$.*

Obviously, for the reverse process, an information system of Boolean-value can be represented as a soft set. As an illustration, let us consider the following example quoted directly after Molodtsov (2004).

Example 1. Let a soft set (F, E) describe the "attractiveness of houses" that Mr. X is going to purchase. Suppose that $U = \{h_1, h_2, h_3, h_4, h_5, h_6\}$ and $E = \{e_1, e_2, e_3, e_4, e_5\}$, where there are six houses in the universe U and E is a set of parameters, $e_i = (i = 1, 2, 3, 4, 5)$ standing for the parameters "expensive", "beautiful", "wooden", "cheap", and "in the green surroundings", respectively.

Suppose that we have $F(e_1) = \{h_2, h_4\}$, $F(e_2) = \{h_1, h_3\}$, $F(e_3) = \phi$, $F(e_4) = \{h_1, h_3, h_5\}$, and $F(e_5) = \{h_1\}$, where $F(e_i)$ means a subset of U whose elements match the parameter e_i. Then we can view the soft set (F, E) as that consisting of the following collection of approximations:

$$(F, E) = \left\{ \begin{array}{r} \text{expensive houses} = \{h_2, h_4\} \\ \text{beautiful houses} = \{h_1, h_3\} \\ \text{wooden houses} = \phi \\ \text{cheap houses} = \{h_1, h_3, h_5\} \\ \text{in the green} \\ \text{surrounding} = \{h_1\} \\ \text{houses} \end{array} \right\}.$$

Each approximation has two parts, a predicate p and an approximate value set v. For example, for the approximation "expensive houses=$\{h_2, h_4\}$", we have the predicate name of expensive houses and the approximate value set or value set is $\{h_2, h_4\}$. Thus, a soft set (F, E) can be viewed as a collection of approximations below:

$$(F, E) = \{p_1 = v_1, p_2 = v_2, p_3 = v_3, \ldots, p_n = v_n\}.$$

♦

The soft set is a mapping from a parameter to the crisp subset of universe. From such a case, we can see that the structure of a soft set can classify the objects into two classes (`yes/1` or `no/0`). Thus we can make a one-to-one correspondence between a Boolean-valued information system and a soft set, as stated in Proposition 1. A soft set (F, E) as in Example 1, can be represented as in Table 1.

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

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

Definition 3. A pair is called an *incomplete soft set* over U if there exists $x_i \in U(i = 1, 2, \ldots, n)$ and $e_j \in E(j = 1, 2, \ldots, m)$, making $x_i \in F(e_j)$ unknown, that is, $F(e_j)(x_i) = null$. In a tabular representation, null is represented by "*".

Example 2. Assume a community college is recruiting some new teachers and there are eight persons applying 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. Suppose several applicants miss 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. Data analysis approaches of soft sets under incomplete information

In this section, we briefly discuss data analysis approaches of soft sets under incomplete information (DASI), which were presented by Zou and Xiao (2008).

Suppose that $U = \{h_1, h_2, \ldots, h_n\}$, $E = \{e_1, e_2, \ldots, e_y\}$, (F, E) is an incomplete soft set with tabular representation. Suppose that for the object h_i there are some incomplete data. Let c_i be the possible choice value of the object h_i, $c_i = \sum_{j=1}^{y} h_{ij}$, where h_{ij} are the entries in the table of (F, E) and y is the number of parameters, c_i be a set of all possible choice values of the object h_i, so the decision value $d_i = \sum_{i=1}^{m} k_i c_i$, where k_i is the weight of the choice value c_i. The weight of the possible choice value is defined as

$$k = \begin{cases} \prod\limits_{e \in E_0^*} q_e, & x = 0, \\ \sum\limits_{C_d^x} [(\prod\limits_{[e_i \in E_1^*]} p_{e_i})(\prod\limits_{[e_j \in E_0^*]} q_{e_j})], & 0 < x < d, \\ \prod\limits_{e \in E_1^*} p_e, & x = d, \end{cases} \tag{1}$$

where d is the number of parameter columns having incomplete information, that is, there are d cells with "*" in the corresponding row in the tabular representation.

For a given choice value of the object h, let x be the number of cells with the value of 1 in this row, and $d - x$ be the number of cells with the value of 0 in this row, E_1^* and E_0^* be the sets of parameters with the value of 1 and 0 for the object h, respectively. Here p_e and q_e stand for probabilities that an object belongs to and does not belong to $F(e)$, respectively, defined by

$$p_e = \frac{n_1}{n_1 + n_0}, \quad q_e = \frac{n_0}{n_1 + n_0}, \quad e \in E, \tag{2}$$

where n_1 and n_0 stand for the number of objects that belong to and do not belong to $F(e)$, respectively.

Based on the above formulas, Zou and Xiao (2008) presented data analysis approaches of soft sets under incomplete information as shown in Fig. 1. From the above approach, we should know that the decision value of an object with incomplete information is calculated by the 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 for the crisp soft sets. It is clear that we only get the decision value for decision making, while the missing data are not still filled. Accordingly the method is only applicable to a decision making system rather than others. It should be found that each choice value of objects is independent of that of other objects. Consequently, the distribution of other available objects deciding the weight of each possible choice value is not reasonable, which makes this method of low accuracy. In order to overcome these problems, we propose a novel data filling approach for incomplete soft sets (DFIS).

4. Novel data filling approach for incomplete soft sets

In this section, we firstly introduce the definition of association degrees between parameters. Furthermore, the related heuristic algorithms are presented based on the association degrees.

4.1. Association degree between parameters. So far, little research has focused on the associations between parameters in 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

(1) Input the incomplete soft set (F, E) and the parameter set E;
(2) Calculate all possible choice values c_i for each object, respectively;
(3) Calculate the weight k_i of the each possible choice value for each object according to formula (1), respectively;
(4) Compute the decision value $d_i = \sum_{d=1}^{m} k_i c_i$ for each object.

Fig. 1. Data analysis approach of soft sets under incomplete information of Zou and Xiao (2008).

know, the attribute "weight" has some relation with the attribute "height". Let us reconsider Examples 1 and 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 and vice versa. There is an inconsistent association between the parameter "expensive" and the parameter "cheap". Generally speaking, if a house is beautiful or is in green surroundings, the house is expensive. There is a consistent association between the parameter "beautiful" and the parameter "expensive" or between "in the green surroundings" and "expensive". Similarly, in Example 2, there is an obvious inconsistent association between the parameter "the highest academic degree is Doctor" and the 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 an inconsistent association between the parameter "experienced" and parameter "young age" or between "married" and "young age".

These associations reveal the interior relations among objects. 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 fill in 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 the association degree and some relative concepts.

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

$$U_{ij} = \{x | F(e_i)(x) \neq \text{`*'} \text{ and } F(e_j)(x) \neq \text{`*'},\ x \in U\}. \tag{3}$$

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 4. Let E be a set of parameters and $e_i, e_j \in E, (i, j = 1, 2, \ldots, m)$. The *consistent association number between parameter* e_i and parameter e_j is denoted by CN_{ij} and defined as

$$CN_{ij} = |\{x | F(e_i)(x) = F(e_j)(x),\ x \in U_{ij}\}|, \tag{4}$$

where m denotes the number of parameters and $|\cdot|$ denotes the cardinality of its set argument.

Example 3. Assume that there exists an incomplete soft set (F, E) with the tabular representation displayed as in Table 2. According to Definition 4, we get that the consistent association number CN_{12} between parameter e_1 and parameter e_2 is 1. Similarly, $CN_{13} = 5$, $CN_{14} = 4$, $CN_{15} = 4$, $CN_{16} = 3$, $CN_{23} = 1$, $CN_{24} = 2$, $CN_{25} = 4$, $CN_{26} = 4$, $CN_{34} = 4$, $CN_{35} = 3$, $CN_{36} == 3$, $CN_{45} = 0$, $CN_{46} = 4$, $CN_{56} = 3$. ♦

Definition 5. Let E be a set of parameters and $e_i, e_j \in E, (i, j = 1, 2, \ldots, m)$. The *consistent association degree* between parameter e_i and parameter e_j is denoted by CD_{ij} and defined as

$$CD_{ij} = \frac{CN_{ij}}{|U_{ij}|}. \tag{5}$$

Obviously, the value of CD_{ij} is in the interval $[0, 1]$. The consistent association degree measures the extent to which the value of parameter e_i is consistent with that of parameter e_j over U_{ij}.

Example 4. Assume that there exists an incomplete soft set (F, E) with the tabular representation displayed as in Table 2. According to Definition 5, we can get the Consistent Association Degree for (F, E) as follows: $CD_{12} = 1/7$, $CD_{13} = 5/7$, $CD_{14} = 4/8$, $CD_{15} = 4/7$, $CD_{16} = 3/7$, $CD_{23} = 1/6$, $CD_{24} = 2/7$, $CD_{25} = 4/6$, $CD_{26} = 4/6$, $CD_{34} = 4/7$, $CD_{35} = 3/6$, $CD_{36} = 3/7$, $CD_{45} = 0/7$, $CD_{46} = 4/7$, $CD_{56} = 3/6$. ♦

Similarly, we can define the inconsistent association number and the inconsistent association degree as follows.

Definition 6. Let E be a set of parameters and $e_i, e_j \in E, (i, j = 1, 2, \ldots, m)$. The *inconsistent association number* between parameter e_i and parameter e_j is denoted by IN_{ij} and is defined as

$$IN_{ij} = |\{x | F(e_i)(x) \neq F(e_j)(x),\ x \in U_{ij}\}|. \tag{6}$$

Definition 7. Let E be a set of parameters and $e_i, e_j \in E(i, j = 1, 2, \ldots, m)$. The *inconsistent association degree* between parameter e_i and parameter e_j is denoted by ID_{ij} and is defined as

$$ID_{ij} = \frac{IN_{ij}}{|U_{ij}|}. \tag{7}$$

Obviously, the value of ID_{ij} is also in $[0, 1]$. The inconsistent association degree measures the extent to which parameters e_i and e_j are inconsistent.

Example 5. Consider incomplete soft set (F, E) with the tabular representation displayed as in Table 2. According to Definition 6, we can get that the consistent association number IN_{12} between parameter e_i and parameter e_j is 6. Similarly, $IN_{13} = 2$, $IN_{14} = 4$, $IN_{15} = 3$, $IN_{16} = 4$, $IN_{23} = 5$, $IN_{24} = 5$, $IN_{25} = 2$, $IN_{26} = 2$, $IN_{34} = 3$, $IN_{35} = 3$, $IN_{36} = 4$, $IN_{45} = 7$, $IN_{46} = 3$, $IN_{56} = 3$. According to Definition 7, naturally, we can get the inconsistent association degree for (F, E) as

follows: $CD_{12} = 6/7$, $CD_{13} = 2/7$, $CD_{14} = 4/8$, $CD_{15} = 3/7$, $CD_{16} = 4/7$, $CD_{23} = 5/6$, $CD_{24} = 5/7$, $CD_{25} = 2/6$, $CD_{26} = 2/6$,$CD_{34} = 3/7$, $CD_{35} = 3/6$, $CD_{36} = 4/7$, $CD_{45} = 7/7$, $CD_{46} = 3/7$, $CD_{56} = 3/6$. ♦

Definition 8. Let E be a set of parameters and $e_i, e_j \in E (i, j = 1, 2, \ldots, m)$. The *association degree* between parameter e_i and parameter e_j is denoted by D_{ij} and defined as

$$D_{ij} = \max \{CD_{ij}, ID_{ij}\}. \tag{8}$$

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

Example 6. Based on Definition 8, association degrees for the incomplete soft set (F, E) are easily obtained as shown in Table 3. Note that "I" means the association degree value is decided by an inconsistent association degree, while "C" means the association degree value is decided by the consistent association degree in Table 3. ♦

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

	e_1	e_2	e_3	e_4	e_5	e_6
e_2	0.86I	–	0.83I	0.71I	0.67C	0.67C
e_3	0.71C	0.83I	–	0.57C	0.5C	0.57I
e_5	0.57C	0.67C	0.5C	1I	–	0.5C
e_6	0.57I	0.67C	0.57I	0.57C	0.5C	–

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

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 greater than 0.5, namely, $D_{ij} = \max \{CD_{ij}, ID_{ij}\} \geq 0.5$. ■

It is obvious that the lowest association degree in Table 3 is 0.5.

Definition 9. Let E be a set of parameters and $e_i, e_j \in E (i, j = 1, 2, \ldots, m)$. The *maximal association degree* of parameter e_i is denoted by D_i and defined as

$$D_i = \max\, D_{ij}, \quad j = 1, 2, \ldots, m. \tag{9}$$

where m is the number of parameters.

4.2. Proposed algorithm for data filling. Below, we provide an algorithm to illustrate how to fill the incomplete data for an incomplete soft set. In our method, first, we calculate association degrees between the parameter e_i and each of the other parameters, respectively, over the existing complete information, and then we 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 the mapping set $F(e_j)$ based on association degrees. However, sometimes a parameter may have a lower maximal association degree, that is, the parameter has a 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 by Zou and Xiao (2008), we can use the probability of objects appearing in $F(e_i)$ to fill the missing data. In any case, in our method we give priority to the association between the parameters instead of the probability of objects appearing in $F(e_i)$ to fill the missing data due to the fact that the relation between the parameters is more reliable than that between the

(1) Input the incomplete soft set (F, E).
(2) Find e_i, which includes missing data $F(e_i)(x)$
(3) Compute $D_{ij}, j = 1, 2, \ldots, m$ where m is the number of parameters in E.
(4) Compute the 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 a consistent association between e_i and e_j, $F(e_i)(x) = F(e_j)(x)$; otherwise there is an 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 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 numbers of objects that belong to and do not belong to $F(e_i)$, respectively.
(8) If $P_1 > P_0$, $F(e_i)(x) = 1$. If $P_1 > P_0$, $F(e_i)(x) = 0$. If $P_1 = P_0$, 0 or 1 may be assigned to $F(e_i)(x)$.
(9) If all of the missing data are filled, the algorithm ends. Otherwise, go to Step 2.

Fig. 2. Proposed algorithm for data filling.

objects in a soft set. Therefore, we can set a threshold λ 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 $F(e_j)$ based on the association degree, or otherwise, the missing data will be filled in terms of the probability of objects appearing in $F(e_i)$. Figure 2 shows the details of the algorithm.

From Figs. 1 and 2, we can observe some differences between DFIS and DASI as follows:

- DASI calculates the decision value of an object with missing data. DFSI calculates the maximal association degree of a parameter with missing data.
- DFSI fills the missing data in an incomplete soft set; however, DASI cannot fill them.
- DASI is only applicable to decision making because it directly calculates the decision value of an object with missing data; however, DFSI converts an incomplete soft set into a complete soft set, which makes the soft set not only applicable to decision making but also useful for other applications.
- The computation of the decision value in DASI completely depends on the probability distribution of other available objects. DFSI gives priority to the association between the parameters instead when filling the missing data, which is more reliable than the distribution of other available objects.

5. Experimental results

In this section, we compare DFIS with DASI. Firstly, a comparison is elaborated through a Boolean data set as shown in Table 2. Both the algorithms are implemented as C++ programs. They are executed sequentially on Intel Core 2 Duo CPUs. Then, the experimental results are analyzed and comparisons are made based on five UCI benchmark datasets. Mainly, we compare two approaches in terms of predictive accuracy. Due to the decision value for decision making in DASI, the predictive accuracy is described by the Mean Absolute Percentage Error (MAPE), which is the most commonly used error measure in business, employed to evaluate forecast models (Hyndman and Koehler, 2006).

5.1. Comparison results for Table 2.

Example 7. Let (F, E) be a soft set with the tabular representation displayed in Table 4, which is the original version of Table 2. Due to the missing of some data, there is an incomplete soft set shown as in Table 2. Suppose that $U = \{c_1, c_2, c_3, c_4, c_5, c_5, c_6, c_7, c_8\}$, and $E = \{e_1, e_2, e_3, e_4, e_5, e_6\}$.

5.1.1. Results from DASI.

Step 1. Calculate all possible choice values c_i for each object having the missing data. For object c_4, the possible choice values are 4, 3, 2; for object c_6, the possible choice values are 2, 1; for object c_7, the possible choice values are 4, 3.

Table 4. Tabular representation of the soft set (F, E) (original version of Table 2).

U	e_1	e_2	e_3	e_4	e_5	e_6	d_i
c_1	1	0	1	0	1	0	3
c_2	1	0	0	1	0	0	2
c_3	0	1	0	0	1	0	2
c_4	0	1	1	1	0	0	3
c_5	1	0	1	1	0	0	3
c_6	0	1	0	0	1	0	2
c_7	1	0	1	0	1	0	3
c_8	0	0	1	1	0	0	2

Step 2. Calculate the weight k_i of the each possible choice value for each object according to the formula (1). For object c_4, there are two cells with "*" in parameter columns e_3 and e_6, respectively. Thus we can get $p_{e_3} = 4/7$,$q_{e_3} = 3/7$, $p_{e_6} = 0$, $q_{e_6} = 1$. According to (1), the weight of the choice value of 4 is $k_1 = p_{e_3}p_{e_6} = 0$, the weight of the choice value of 3 is $k_2 = p_{e_3}q_{e_6} + q_{e_3}p_{e_6} = 4/7$, and the weight of the choice value of 2 is $k_3 = q_{e_3}q_{e_6} = 3/7$. For the object c_6, there is one cell with "*" in the parameter column e_5. So we can get $p_{e_5} = 3/7$, $q_{e_5} = 4/7$. According to (1), the weight of the choice value of 2 is $k_1 = p_{e_5} = 3/7$, the weight of the choice value of 1 is $k_2 = q_{e_5} = 4/7$. For the object e_7, there is one cell with "*" in the parameter column e_2. Thus we can get $p_{e_2} = 3/7$,$q_{e_2} = 4/7$. According (1), the weight of the choice value of 4 is $k_1 = p_{e_5} = 3/7$, the weight of the choice value of 3 is $k_2 = q_{e_5} = 4/7$.

Step 3. Compute the decision value $d_i = \sum_{i=1}^{m} k_i c_i$ for each object. Therefore, $d_4 = k_1 \times 4 + k_2 \times 3 + k_3 \times 2 = 18/7$; similarly, $d_6 = k_1 \times 2 + k_2 \times 1 = 10/7$, $d_7 = k_1 \times 4 + k_2 \times 3 = 24/7$.

5.1.2. Results from DFIS.

Suppose that the threshold is $\lambda = 0.8$.

Step 1. Find the missing data. Here $F(e_2)(c_7)$, $F(e_3)(c_4)$, $F(e_5)(c_6)$ and $F(e_6)(c_4)$ need to be filled.

Step 2. Compute the association degrees $D_{ij}, j = 1, 2, \ldots, m$, which is shown as in Table 3.

Step 3. Compute the maximal association degree D_i for missing data, if $D_i \geq \lambda$, and then fill them according to

the association degree. For parameter e_2, from Table 3 we can see 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 from the inconsistent association degree, D_{25} and D_{26} are from the consistent association degree. The maximal association degree $D_2 = 0.86$. We set the threshold $\lambda = 0.8$. Therefore, in terms of the proposed algorithm, we can fill $F(e_2)(e_7)$ according to $F(e_1)(e_7)$. Because $F(e_2)(e_7) = 1$ and there is an inconsistent association between parameters e_2 and e_1, we fill 0 into $F(e_2)(e_7)$. Similarly, we can fill 0, 1 into $F(e_3)(e_4)$ and $F(e_5)(e_6)$, respectively.

Step 4. If $D_i < \lambda$, compute the probabilities P_1 and P_0 and fill the missing data according to the probabilities. For parameter e_6, we have the maximal association degree $D_6 = 0.67 < \lambda$. That means there is no reliable association between parameter e_6 and other parameters. So we can fill the data $F(e_6)(c_4)$ according to the probabilities of other available objects. Here, we have $P_0 = 1$,$P_1 = 0$. Therefore, we fill 0 into $F(e_6)(c_4)$. Table 5 shows the tabular representation of the filled soft set of Table 2.

Table 5. Tabular representation of the filled soft set of Table 2.

U	e_1	e_2	e_3	e_4	e_5	e_6	d_i
c_1	1	0	1	0	1	0	3
c_2	1	0	0	1	0	0	2
c_3	0	1	0	0	1	0	2
c_4	0	1	0	1	0	0	2
c_5	1	0	1	1	0	0	3
c_6	0	1	0	0	1	0	2
c_7	1	0	1	0	1	0	3
c_8	0	0	1	1	0	0	2

We compare two approaches in terms of the predictive accuracy, which is described by the mean absolute percentage error. The mean absolute percentage error is defined as

$$M = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{A_t - F_t}{A_t}\right|, \tag{10}$$

where A_t is the actual decision value and F_t is the forecast decision value.

We denote by M_1 the MAPE of DASI and by M_2 the MAPE of DFIS. Thus,

$$M_1 = \frac{1}{3}\left(\left|\frac{3-18/7}{3}\right| + \left|\frac{2-10/7}{2}\right| + \left|\frac{3-24/7}{3}\right|\right) = 0.1905.$$

Similarly, $M_2 = 0.1111$. It is obvious that our approach has a smaller MAPE compared with DASI, and thus it improves the forecasting accuracy. ♦

5.2. Comparison results for UCI benchmark datasets. This section discusses the experiment that was made using DFIS and DASI on five UCI benchmark datasets (UCI, 2012), i.e, Zoo, SPECT Heart, Congressional Votes, Acute Inflammations, and Flag. Here the given association degree threshold $\lambda = 0.8$. Since the soft set is a Boolean-valued information system, we only need Boolean parameters to be tested for these datasets. On each dataset, we randomly delete 10 entries, 20 entries, 30 entries, 40 entries, 50 entries, 60 entries, 70 entries and 80 entries as the missing data, respectively. Given the number of missing data, we run the program 100 times and compute the average MAPE as the final one. Running DFIS and DASI with a different number of missing data, two vectors of MAPE finally are obtained on each dataset. To determine whether the means of the two vectors are statistically different from each other, a t-test is also performed. The t-test is implemented by using the function $h = ttest2(x, y)$ in Matlab Statistics Toolbox, where x and y denote the vectors of the MAPE obtained by DFIS and DASI, respectively. The result $h = 1$ means that we can reject the hypothesis that the means are equal at the 0.05 significance level and $h = 0$ otherwise. Experimental results are shown in what follows.

5.2.1. Zoo dataset. The Zoo dataset consists of 101 instances of animals with 18 features. The names of the animals constitute the first attribute. There are 15 Boolean features in terms of the presence of hair, feathers, backbone, eggs, fins, eggs, tail, and of whether the animals are airborne, aquatic, predator, toothed, catsize, domestic, breathes, venomous. Since the soft set is a Boolean-valued information system, we only need Boolean parameters to be tested. Using DFIS and DASI, we get two groups of MAPE values when the number of missing data varies from 10 to 80, namely, vectors x and y. Here

$$x = [0.036, 0.032, 0.033, 0.034, 0.033, 0.031, 0.057, 0.053],$$

$$y = [0.069, 0.068, 0.069, 0.069, 0.07, 0.068, 0.063, 0.059].$$

Figure 3 shows the MAPE values obtained by two approaches to the Zoo dataset. It is obvious that our approach has a lower MAPE value compared with DASI, and thus it improves the forecasting accuracy. DFSI improves the forecasting accuracy of DASI up to on 42.2% average. Especially, when the number of missing values ranges from 10 to 60, DFSI improves the forecasting accuracy of DASI up to 52.2% on the average. Further, a t-test is performed on vectors x and y, namely, $h = ttest2(x, y)$. The result is $h = 1$, which indicates the means of vectors x and y are statistically different. The

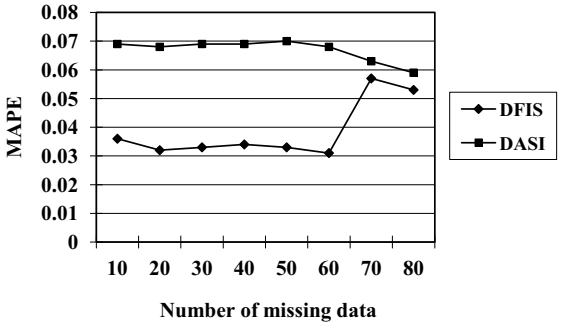

Fig. 3. MAPE of two approaches for the Zoo dataset.

mean of x is less than that of y, that is, the forecasting accuracy of DFIS is higher than that of DASI on the Zoo dataset.

5.2.2. SPECT Heart dataset. The dataset describes the diagnosing of cardiac Single Proton Emission Computed Tomography (SPECT) images. Here we use the SPECT Heart training data set which contains 80 patients with 22 categorical attributes. This dataset is originally a Boolean dataset. Using DFIS and DASI, we get MAPE vectors x and y. Here

$$x = [0.038, 0.031, 0.034, 0.032, 0.035, 0.036, 0.033, 0.034],$$

$$y = [0.095, 0.095, 0.094, 0.095, 0.094, 0.094, 0.094, 0.094].$$

Figure 4 illustrates the MAPE values obtained by two approaches to the SPECT Heart dataset.

In this case, DFIS improves the forecasting accuracy of DASI up to 63.8% on the average. Further, a t-test $h = ttest2(x, y)$ is performed on vectors x and y. The result is $h = 1$, which indicates the means of vectors x and y are statistically different. The mean of x is less than that of y, that is, the forecasting accuracy of DFIS is higher than that of DASI for the SPECT Heart dataset.

5.2.3. Congressional Votes. This example concerns the United States Congressional Voting Records in 1984. Each record represents one congressman's votes regarding 16 issues. All attributes are Boolean with *yes* and *no* values. Originally, the dataset contains 435 objects, including 203 objects having the missing data. In order to reveal the MAPE of two approaches, we only reserve 232 objects having complete data. Using DFIS and DASI,

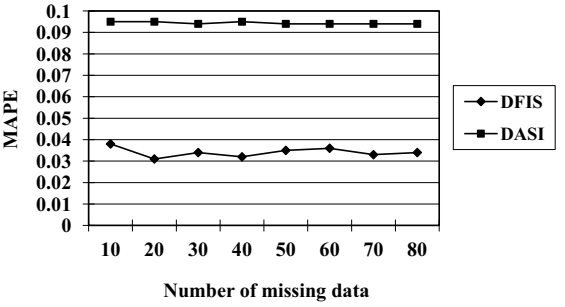

Fig. 4. MAPE of two approaches for the SPECT Heart dataset.

we get MAPE vectors x and y. Here

$$x = [0.026, 0.027, 0.026, 0.026, 0.025, 0.027, 0.026, 0.027],$$

$$y = [0.059, 0.059, 0.059, 0.059, 0.059, 0.059, 0.059, 0.059].$$

Figure 5 illustrates the MAPE values obtained by the two approaches for the Congressional Votes dataset. In this case, DFIS improves the forecasting accuracy of

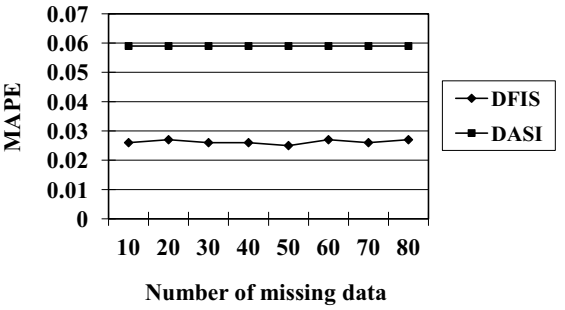

Fig. 5. MAPE of two approaches for the Congressional Votes dataset.

DASI up to 55.9% on the average. Further, a t-test $h = ttest2(x, y)$ is performed on vectors x and y. The result is $h = 1$, which indicates the means of vectors x and y are statistically different. The mean of x is less than that of y, that is, the forecasting accuracy of DFIS is higher than that of DASI on the Congressional Votes dataset.

5.2.4. Acute Inflammations dataset. This dataset was created by a medical expert as a data set to test the expert system which will perform the presumptive diagnosis of two diseases of the urinary system. There are 120 instances and 6 parameters. However, one of the

parameters is non-Boolean. So we choose 5 parameters to be tested. Using DFIS and DASI, we get MAPE vectors x and y. Here

$$x = [0.174, 0.173, 0.175, 0.178, 0.183, 0.174, 0.178, 0.173],$$

$$y = [0.237, 0.227, 0.231, 0.229, 0.232, 0.23, 0.232, 0.23].$$

Figure 6 illustrates the MAPE values obtained by the two approaches for the Acute Inflammations dataset. In this case, DFIS improves the forecasting accuracy of DASI up to 23.8% on the average. Further, a t-test $h = ttest2(x, y)$ is performed on vectors x and y. The result is $h = 1$, which indicates the means of vectors x and y are statistically different. The mean of x is less than that of y, that is, the forecasting accuracy of DFIS is higher than that of DASI for the Acute Inflammations dataset.

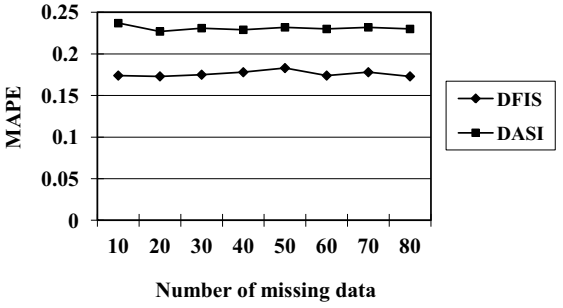

Fig. 6. MAPE of two approaches for the Acute Inflammations dataset.

5.2.5. Flag database. This dataset contains details of various nations and their flags. Originally, there are 194 instances and 28 parameters. Due to a Boolean information system, 12 parameters are chosen to be tested. Using DFIS and DASI, we get MAPE vectors x and y. Here

$$x = [0.062, 0.065, 0.066, 0.069, 0.067, 0.069, 0.065, 0.068],$$

$$y = [0.095, 0.092, 0.092, 0.094, 0.093, 0.093, 0.092, 0.093].$$

Figure 7 illustrates the MAPE values obtained by the two approaches for the Flag dataset. In this case, DFIS improves the forecasting accuracy of DASI up to 29.0% on the average. Further, a t-test $h = ttest2(x, y)$ is

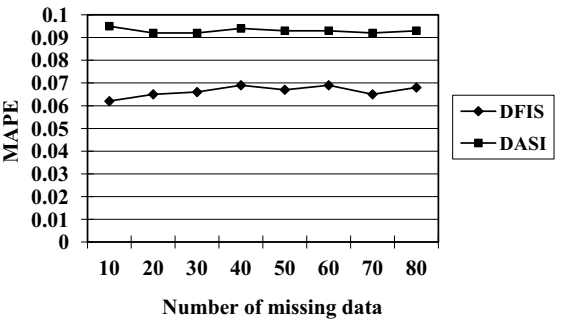

Fig. 7. MAPE of two approaches for the Flag dataset.

performed on vectors x and y. The result is $h = 1$, which indicates the means of vectors x and y are statistically different. The mean of x is less than that of y, that is, the forecasting accuracy of DFIS is higher than that of DASI for the Flag dataset. We summarize the aforementioned experimental results as follows:

- DFIS improves the forecasting accuracy of DASI on all five UCI datasets. The forecasting accuracy of DASI is improved up to 42.9% on the average for these five datasets.
- The results of the t-test on five datasets are the same, namely, $h = 1$. This result shows that the means of MAPE vectors obtained by DFIS and DASI are statistically different, which provides a statistical rigor to the improved accuracy by DFIS.

6. Conclusions

Data filling for an incomplete soft set is rarely studied in the soft set literature. The paper aims to introduce a novel data filling approach for incomplete soft sets (DFIS), which is based on the association degree among parameters. DFIS can overcome the two inherent deficiencies in DASI. Data filling converts an incomplete soft set into a complete soft set, which makes the soft set not only applicable to decision making but also useful for other applications, while the approach of DASI is only applicable to decision making. The comparison results elaborated between the two approaches through five UCI benchmark datasets illustrate that DFIS outperforms the approach of DASI regarding the forecasting accuracy.

References

Aktas, H. and Cagman, N. (2007). Soft sets and soft groups, *Information Sciences* **177**(13): 2726–2735.

Ali, M., Feng, F., Liu, X., Min, W. and Shabira, M. (2009). On some new operations in soft set theory, *Computers and Mathematics with Applications* **57**(9): 1547–1553.

Atanassov, K. (1986). Intuitionistic fuzzy sets, *Fuzzy Sets and Systems* **20**(1): 87–96.

Çağman, N. and Enginoğlu, S. (2010). Soft matrix theory and its decision making, *Computers and Mathematics with Applications* **59**(10): 3308–3314.

Chen, D., Tsang, E., Yeung, D. and Wang, X. (2005). The parameterization reduction of soft sets and its applications, *Computers and Mathematics with Applications* **49**(5–6): 757–763.

Feng, F. (2009). Generalized rough fuzzy sets based on soft sets, *Proceedings of the 2009 International Workshop on Intelligent Systems and Applications, ISA 2009, Wuhan, China,* pp. 1–4.

Feng, F., Jun, Y., Liu, X. and Li, L. (2010a). An adjustable approach to fuzzy soft set based decision making, *Journal of Computational and Applied Mathematics* **234**(1): 10–20.

Feng, F., Li, Y. and Leoreanu-Fotea, V. (2010b). Application of level soft sets in decision making based on interval-valued fuzzy soft sets, *Computers and Mathematics with Applications* **60**(6): 1756–1767.

Feng, F., Jun, Y. and Zhao, X. (2008). Soft semirings, *Computers and Mathematics with Applications* **56**(10): 2621–2628.

Feng, F., Li, C., Davvaz, B. and Ali, M. (2009). Soft sets combined with fuzzy sets and rough sets: A tentative approach, *Soft Computing* **14**(9): 899–911.

Gau, W. and Buehrer, D. (1993). Vague sets, *IEEE Transactions on System, Man, and Cybernetics* **23**(2): 610–614.

Gorzalzany, M. (1987). A method of inference in approximate reasoning based on interval-valued fuzzy sets, *Fuzzy Sets and Systems* **21**(1): 1–17.

Herawan, T. and Mat Deris, M. (2009). A direct proof of every rough set is a soft set, *Proceedings of the 3rd Asia International Conference on Modeling and Simulation, AMS'09, Bali, Indonesia,* pp. 119–124.

Herawan, T. and Mat Deris, M. (2011). A soft set approach for association rules mining, *Knowledge Based Systems* **24**(1): 186–195.

UCI (2012). Benchmark datasets, `http://www.ics.uci.edu/mlearn/`.

Hyndman, R. and Koehler, A. (2006). Another look at measures of forecast accuracy, *International Journal of Forecasting* **22**(4): 679–688.

Jiang, Y., Tang, Y., Chen, Q., Liu, H. and Tang, J. (2010). Interval-valued intuitionistic fuzzy soft sets and their properties, *Computers and Mathematics with Applications* **60**(3): 906–918.

Jiang, Y., Tang, Y., and Chen, Q. (2011). An adjustable approach to intuitionistic fuzzy soft sets based decision making, *Applied Mathematical Modelling* **35**(2): 824–836.

Jun, Y. and Park, C. (2008). Applications of soft sets in ideal theory of BCK/BCI-algebras, *Information Sciences* **178**(11): 2466–2475.

Kong, Z., Gao, L. and Wang, L. (2009). Comment on "A fuzzy soft set theoretic approach to decision making problems", *Journal of Computational and Applied Mathematics* **223**(2): 540–542.

Kong, Z., Gao, L., Wang, L. and Li, S. (2008). The normal parameter reduction of soft sets and its algorithm, *Computers and Mathematics with Applications* **56**(12): 3029–3037.

Li, C. and Chiang, T.-W. (2011). Function approximation with complex neuro-fuzzy system using complex fuzzy sets—A new approach, *New Generation Computing* **29**(3): 261–276.

Ma, X., Sulaiman, N., Qin, H., Herawan, T. and Zain, J. (2011). A new efficient normal parameter reduction algorithm of soft sets, *Computers and Mathematics with Applications* **62**(2): 588–598.

Maji, P. (2009). More on intuitionistic fuzzy soft sets, *in* H. Sakai, M. Chakraborty, A. Hassanien, D. Slezak and W. Zhu (Eds.), *Proceedings of the 12th International Conference on Rough Sets, Fuzzy Sets, Data Mining and Granular Computing (RSFDGrC 2009)*, Lecture Notes in Computer Science, Vol. 5908, Springer, Berlin/Heidelberg, pp. 231–240.

Maji, P., Biswas, R. and Roy, A. (2001a). Fuzzy soft sets, *Journal of Fuzzy Mathematics* **9**(3): 589–602.

Maji, P., Biswas, R. and Roy, A. (2001b). Intuitionistic fuzzy soft sets, *Journal of Fuzzy Mathematics* **9**(3): 677–692.

Maji, P., Biswas, R. and Roy, A. (2003). Soft set theory, *Computers and Mathematics with Applications* **45**(4–5): 555–562.

Maji, P. and Roy, A. (2002). An application of soft sets in a decision making problem, *Computers and Mathematics with Applications* **44**(8–9): 1077–1083.

Maji, P. and Roy, A. (2007). A fuzzy soft set theoretic approach to decision making problems, *Journal of Computational and Applied Mathematics* **203**(2): 412–418.

Maji, P., Roy, A. and Biswas, R. (2004). On intuitionistic fuzzy soft sets, *Journal of Fuzzy Mathematics* **12**(3): 669–683.

Majumdar, P. and Samanta, S. (2010). Generalized fuzzy soft sets, *Computers and Mathematics with Applications* **59**(4): 1425–1432.

Molodtsov, D. (1999). Soft set theory first results, *Computers and Mathematics with Applications* **37**(4–5): 19–31.

Molodtsov, D. (2004). *The Theory of Soft Sets*, URSS Publishers, Moscow, (in Russian).

Nowicki, R. (2010). On classification with missing data using rough-neuro-fuzzy systems, *International Journal of Applied Mathematics and Computer Science* **20**(1): 55–67, DOI: 10.2478/v10006-010-0004-8.

Pawlak, Z. (1982). Rough sets, *International Journal of Computing and Information Sciences* **11**(5): 341–356.

Pei, D. and Miao, D. (2005). From soft sets to information systems, *Proceedings of the 2005 IEEE International Conference on Granular Computing, IEEE GrC'05, Beijing, China*, pp. 617–621.

Qin, H., Ma, X., Herawan, T. and Zain, J. (2011a). An adjustable approach to interval-valued intuitionistic fuzzy soft sets based decision making, *in* N. Nguyen, C. Kim and A. Janiak (Eds.), *ACIIDS 2011*, Lecture Notes in Computer Science, Vol. 6592, Springer, Berlin/Heidelberg, pp. 80–89.

Qin, H., Ma, X., Herawan, T. and Zain, J. (2011b). Data filling approach of soft sets under incomplete information, *in* N. Nguyen, C.G. Kim and A. Janiak (Eds.), *ACIIDS 2011*, Lecture Notes in Computer Science, Vol. 6592, Springer, Berlin/Heidelberg, pp. 302–311.

Qin, K. and Kong, Z. (2010). On soft equality, *Journal of Computational and Applied Mathematics* **234**(5,1): 1347–1355.

Quinlan, J. (1989). Unknown attribute values in induction, *Proceedings of the 6th International Machine Learning Workshop, San Mateo, Canada*, pp. 164–168.

Thiesson, B. (1995). Accelerated quantification of Bayesian networks with incomplete data, *1st International Conference on Knowledge Discovery and Data Mining, Montreal, Canada*, pp. 306–311.

Xiao, Z., Gong, K. and Zou, Y. (2009). A combined forecasting approach based on fuzzy soft sets, *Journal of Computational and Applied Mathematics* **228**(1): 326–333.

Xu, W., Ma, J., Wang, S. and Hao, G. (2010). Vague soft sets and their properties, *Computers and Mathematics with Applications* **59**(2): 787–794.

Yang, X., Lin, T., Yang, J. and Dongjun, Y. (2009). Combination of interval-valued fuzzy set and soft set, *Computers and Mathematics with Applications* **58**(3): 521–527.

Zadeh, L. (1965). Fuzzy sets, *Information Control* **8**(3): 338–353.

Zhang, D. and Li, X. (2006). An absolute information quantity-based data making-up algorithms of incomplete information system, *Computer Engineering and Applications* **42**(22): 155–157.

Zhong, N. and Skowron, A. (2001). A rough set-based knowledge discovery process, *International Journal of Applied Mathematics and Computer Science* **11**(3): 603–619.

Zou, Y. and Xiao, Z. (2008). Data analysis approaches of soft sets under incomplete information, *Knowledge Based System* **21**(8): 941–945.

Hongwu Qin has been a lecturer at Northwest Normal University since 1999. He received the M.Sc. degree in computer science from the Beijing University of Technology in 2005. He is currently a Ph.D. candidate at the Faculty of Computer Systems and Software Engineering, University of Malaysia, Pahang. He has published over 20 journal and conference papers. His research interests include soft sets, KDD and data mining.

Xiuqin Ma is currently a Ph.D. candidate at the Faculty of Computer Systems and Software Engineering, University of Malaysia, Pahang. She has published over 20 journal and conference papers. Her research interests include soft sets, rough sets, and data mining.

Tutut Herawan received his B.Ed. and M.Sc. degrees in mathematics from Ahmad Dahlan University and Gadjah University, Mada Yogyakarta, Indonesia, respectively. He obtained his Ph.D. from Tun Hussein Onn, Malaysia. Currently, he is a senior lecturer at the Department of Computer Science and the leader of the Database and Knowledge Management (DBKM) research group at the Faculty of Computer Systems and Software Engineering, University of Malaysia, Pahang. He has published more than 70 research papers in journals and conference proceedings. His research area includes data mining and KDD as well as rough and soft set theories.

Jasni Mohamad Zain received her Bachelor degree in computer science from the University of Liverpool, UK, in 1989; PGCE in mathematics from Sheffield Hallam University, UK in 1994; M.Ed. from Hull University, UK, in 1998, and Ph.D. from Brunel University, London, UK, in 2005. She currently holds the post of the dean of the Faculty of Computer Systems and Software Engineering, University of Malaysia, Pahang. Her research interests include image processing, data mining as well as data and network security.

Received: 18 October 2011
Revised: 18 May 2012