Knowledge-Based Systems 36 (2012) 139–145

Contents lists available at SciVerse ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

A novel soft set approach in selecting clustering attribute ☆

Hongwu Qin [a,b], Xiuqin Ma [a,b], Jasni Mohamad Zain [a], Tutut Herawan [a,*]

[a] Faculty of Computer Systems and Software Engineering, Universiti Malaysia Pahang, Lebuhraya Tun Razak, Gambang 26300, Kuantan, Malaysia
[b] College of Mathematics and Information Science, Northwest Normal University, Lanzhou Gansu 730070, China

ARTICLE INFO

Article history:
Received 31 December 2011
Received in revised form 17 May 2012
Accepted 2 June 2012
Available online 15 June 2012

Keywords:
Soft set
Rough set
Information system
Clustering attribute

ABSTRACT

Clustering is one of the most useful tasks in data mining process for discovering groups and identifying interesting distributions and patterns in the underlying data. One of the techniques of data clustering was performed by introducing a clustering attribute. Soft set theory, initiated by Molodtsov in 1999, is a new general mathematical tool for dealing with uncertainties. In this paper, we define a soft set model on the equivalence classes of an information system, which can be easily applied in obtaining approximate sets of rough sets. Furthermore, we use it to select a clustering attribute for categorical datasets and a heuristic algorithm is presented. Experiment results on fifteen UCI benchmark datasets showed that the proposed approach provides a faster decision in selecting a clustering attribute as compared with maximum dependency attributes (MDAs) approach up to 14.84%. Furthermore, MDA and NSS have a good scalability i.e. the executing time of both algorithms tends to increase linearly as the number of instances and attributes are increased, respectively.

1. Introduction

Clustering is one of the most useful tasks in data mining process for discovering groups and identifying interesting distributions and patterns in the underlying data, which is required in a number of data analysis tasks, such as unsupervised classification and data summation, as well as segmentation of large homogeneous data sets into smaller homogeneous subsets that can be easily managed, separately modeled and analyzed [1]. Rough set theory [2–4], introduced by Z. Pawlak in 1982, is a mathematical tool to deal with vagueness and uncertainty. It has been widely used in many branches of artificial intelligence and data mining [5–14]. One of the popular techniques in data clustering is based on rough set theory. The main approach of rough set-based data clustering is the clustering dataset which is mapped as the decision table and this approach can be performed by introducing a clustering attribute [15]. Hence, for many candidates in a database, selecting only one attribute, which is the best way to partition the objects, is of primary importance for this approach. Currently, there have been works in the area of applying rough set theory in the process of selecting clustering attribute.

Mazlack et al. [16] proposed two techniques to select clustering attribute, that is , Bi-Clustering (BC) based on balanced/unbalanced bi-valued attributes and Total Roughness (TR) based on the average of the accuracy of approximation (accuracy of roughness) in the rough set theory. Parmar et al. [17] proposed a strategy called Min–Min Roughness (MMR) for categorical data clustering. In selecting clustering attribute, the accuracy of approximation is measured using the well-known Marczeweski–Steinhaus metric applied to the lower and upper approximations of a subset of the universe in the information system [18]. Herawan et al. [19] proposed a new method called maximum dependency of attributes (MDAs) in selecting clustering attribute, which is based on the rough set theory by taking into account the dependency of attributes of the database. Compared to TR and MMR, the MDA provides a better performance. However, these algorithms for categorical data clustering are relatively new, with a strong focus on evaluating performance. In reviewing TR, MMR, and MDA for handling large datasets, the ever-increasing computing capabilities, and computation complexity are still an outstanding issue.

In 1999, Molodtsov [20] proposed a soft set theory as a new mathematical tool for dealing with vagueness and uncertainties. The soft set theory, including the theory of probability and fuzzy sets [21], is different from traditional tools for dealing with uncertainties, and further it is free from the inadequacy of the parameterization tools of those theories [22]. At present, work on the soft set theory is progressing rapidly both in theoretical models and applications. Recently, the relation between the rough set and soft set has also attracted much attention. Feng et al. [23,24] investigated the problem of combining soft with fuzzy and rough

☆ An early version of this paper appeared in the Proceeding of the 2nd International Conference on Software Engineering and Computer System (ICSECS) 2011, Kuantan, Pahang, Malaysia, June 27–29, 2011, *Communications in Computer and Information Science*, Volume 180, Part 2, 16–27, Springer-Verlag Berlin Heidelberg, 2011.

* Corresponding author. Tel.: +60 142723760.
E-mail addresses: qhwump@gmail.com (H. Qin), xueener@gmail.com (X. Ma), jasni@ump.edu.my (J.M. Zain), tutut@ump.edu.my (T. Herawan).

0950-7051/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.knosys.2012.06.001

sets. Three different types of hybrid models were presented, which were called rough soft sets, soft rough sets and soft-rough fuzzy sets respectively. Herawan and Mat Deris gave a direct proof in [25] that every rough set is a soft set. As for practical applications of soft set theory, great progress has been achieved. Soft set theory can be applied to solving the decision-making problem [26–29], data analysis under incomplete information [30], the combined forecasting [31,32], and association rules mining [33].

Here, we have summarized four main contributions to solve the above-mentioned rough set-based technique problems:

(a) We propose a new soft set model for the information system. It is constructed over the set of equivalence class instead of single object.
(b) We apply the proposed soft set model to obtain approximation rough set.
(c) We use the proposed soft set model to select clustering attribute for categorical datasets and a heuristic algorithm is presented.
(d) We elaborate the proposed soft set model on experiment tests through fifteen UCI benchmark data sets and the results revealed that the proposed approach provides a faster decision in selecting a clustering attribute as compared with Maximum Dependency Attributes (MDAs) approach.

The rest of this paper is structured as follows. Section 2 briefly reviews some basic definitions in rough set and soft set theory. The next section describes the analysis of MDA approach. Section 4 depicts the construction of the soft set model on information system and shows the application of the proposed model in selecting clustering attributes. This is followed by a comparison of results between MDA and the proposed technique. Finally, the conclusion of this piece of research is given.

2. Essential rudiments

2.1. Rough set theory

Motivation for rough set theory is needed to represent a subset of a universe in terms of equivalence classes of a partition of the universe. In this section, the basic concept of rough set theory is presented. The notion of information system provides a convenient tool for the representation of objects in terms of their attribute values. An *information system* is a 4-tuple (quadruple) $S = (U, A, V, f)$, where $U = \{u_1, u_2, \cdots, u_{|U|}\}$ is a non-empty finite set of objects, $A = \{a_1, a_2, \cdots, a_{|A|}\}$ is a non-empty finite set of attributes, $V = \bigcup_{a \in A} V_a$, V_a is the domain (value set) of 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 information (knowledge) function [4]. Two elements $x, y \in U$ in S is said to be *B-indiscernible* (indiscernible by the set of attribute $B \subseteq A$ in S) if and only if $f(x, a) = f(y, a)$, for every $a \in B$ [4]. An indiscernible relation induced by the set of attribute B, denoted by $IND(B)$, is an equivalence relation. It is well-known that an equivalence relation can induce a unique partition. The partition of U induced by $IND(B)$ in $S = (U, A, V, f)$ denoted by U/B and the equivalence class in the partition U/B contains $x \in U$ and denotes by $[x]_B$. Let B be any subset of A in S and let X be any subset of U, the *B-lower approximation* of X, denoted by $\underline{B}(X)$ and *B-upper approximation* of X, denoted by $\overline{B}(X)$ respectively, are defined by $\underline{B}(X) = \{x \in U | [x]_B \subseteq X\}$ and $\overline{B}(X) = \{x \in U \mid [x]_B \cap X \neq \phi\}$. The *accuracy of approximation* of any subset $X \subseteq U$ with respect to$B \subseteq A$, denoted by $\alpha_B(X)$ is measured by $\alpha_B(X) = |\underline{B}(X)| / |\overline{B}(X)|$, where $|X|$ denotes the cardinality of X. For empty set ϕ, it is defined that $\alpha_B(\phi) = 1$ [4]. Obviously, $0 \leqslant \alpha_B(X) \leqslant 1$. If X is a union of some equivalence classes of U, then $\alpha_B(X) = 1$. Thus, the set X is *crisp* (precise) with respect to B. And, if X is not a union of some equivalence classes of U, then $\alpha_B(X) < 1$. Thus, the set X is *rough* (imprecise) with respect to B. This means that the higher the accuracy of approximation of any subset $X \subseteq U$, the more precise (the less imprecise) of itself [4].

2.2. Soft set theory

In 1999, Molodtsov [20] proposed the soft set theory claiming that this is a new mathematical tool for dealing with vagueness and uncertainties. A soft set is a parameterized family of the subsets of a universal set. In this sub-section, we present the notion, an example, and a property of this theory. 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. The definition of soft set is given below:.

Definition 1. (See [20]). A pair (F, E) is called a soft set over U, where F is a mapping given by $F: E \to P(U)$.

That is, a soft set over U is a parameterized family of subsets of the universe U. As an illustration, let us consider the following example, which is quoted directly from [20].

Example 1. A soft set (F, E) describes 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 five houses in the universe whereby U and E is a set of parameters, e_i $(i = 1, 2, 3, 4, 5)$ representing the parameters "expensive", "beautiful", "wooden", "cheap", and "in the green surroundings" respectively. Suppose $F(e_1) = \{h_2, h_4\}$, $F(e_2) = \{h_1, h_3\}$, $F(e_3) = \phi$, $F(e_4) = \{h_1, h_3, h_5\}$, and $F(e_5) = \{h_1\}$, where $F(e_i)$ means a subset of $U$with elements match with parameter e_i. Then we can view the soft set (F, E) as a collection of approximations as below:

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

Each approximation has two parts, a predicate p and an approximate value set v. For instance, in the example of the approximation "expensive houses ={h_2, h_4}", the predicate name of expensive houses and the approximate value set is {h_2, h_4}.

Thus, a soft set (F, E) can be viewed as a collection of approximations:

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

The soft set is a mapping from parameter to the crisp subset of universe. From such case, 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.

Proposition 1. (See [13]). If (F, E) is a soft set over the universe U, then (F, E) is a Boolean-valued information system $S = (U, A, V_{\{0,1\}}, f)$.

According to Proposition 1, the soft set (F, E) mentioned in Example 1 can be represented as a Boolean table, which is shown in Table 1:

3. Rough set-based technique for clustering attribute selection

Rough set-based clustering attribute selection approaches for categorical data have attracted much attention in recent years. Herawan et al. [19] proposed a new technique called Maximum

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

U/E	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	1	1	0
h_4	1	0	1	0	0
h_5	0	0	1	1	0
h_6	0	0	0	0	0

Dependency Attributes (MDA) in selecting clustering attribute, which is based on rough set theory by taking into account the dependency of attributes of the database. Compared to TR [16] and MMR [17], MDA technique provides a better performance in terms of executing time. Further, in this section, we briefly introduce and analyze the MDA technique.

3.1. Rough set-based attributes dependency

Discovering dependencies between attributes is one of the important issues in database analysis. Intuitively, a set of attributes D depends totally on a set of attributes C, denoted $C \Rightarrow D$, if all values of attributes from D are uniquely determined by values of attributes from C. In other words, D depends totally on C, if there is a functional dependency between the values of D and C. Since the concepts in information systems are generalizations of the same concept in relational databases, there is a need to understand the notion of functional dependency of attributes, called a *partial dependency*of attributes. The idea of generalized dependency of attributes is given in the following definition:

Definition 2. Let a_i and a_j be the two attributes in the information system $S = (U, A, V, f)$. Attribute a_i depends on a_j in a degree of $k_{a_j}(a_i)$, which is obtained as follows:

$$k_{a_j}(a_i) = \frac{\sum_{X \in U/a_i} |\underline{a_j}(X)|}{|U|}. \quad (1)$$

Obviously, $0 \leqslant k_{a_j}(a_i) \leqslant 1$. If all set X are crisp, then $k_{a_j}(a_i) = 1$. The formula $\sum_{X \in U/a_i} |\underline{a_j}(X)|$, a lower approximation of the partition U/a_i with respect to a_j, is the set of all elements of U that can be uniquely classified to blocks of the partition U/a_i, by means of a_j. Attribute a_i is said to be *fully dependent* (in a degree of $k_{a_j}(a_i)$) on a_j if $k_{a_j}(a_i) = 1$. Otherwise, a_i depends partially on a_j. If all elements of the universe U can be uniquely classified into equivalence classes of the partition U/a_i, then we have to employ a_j.

Example 2. A small dataset describing objects' appearance is shown in Table 2.

From the table, we have three partitions induced by an indiscernible relation of each attribute, i.e.

$$U/\text{Shape} = \{\{1,2\},\{3,4\},\{5\}\}, \quad U/\text{Color} = \{\{1,2\},\{3,5\},\{4\}\},$$
$$U/\text{Area} = \{\{1\},\{2,3,4,5\}\} \quad (2)$$

Table 2
An information system of objects' appearance.

U/A	Shape	Color	Area
1	Circle	Red	Big
2	Circle	Red	Small
3	Triangle	Blue	Small
4	Triangle	Green	Small
5	Circle	Blue	Small

From the partition, there is a fully dependency of attribute area on attribute 'Color', i.e. $\{\text{Color}\} \Rightarrow_{k=1} \{\text{Area}\}$, where k is calculated as follows:

$$k_{\text{Color}}(\text{Area}) = \frac{\sum_{X \in U/\text{Color}=} |\underline{\text{Color}}(X)|}{|U|} = \frac{|\{1,2\}| + |\{3,5\}| + |\{4\}|}{|1,2,3,4,5|} = 1.$$

For dependency attribute $\{\text{Color}\} \Rightarrow_k \{\text{Area}\}$, the degree value is given by $k = \frac{2}{5}$, because the area of two objects can be uniquely determined by employing the attribute 'Shape'.

3.2. Maximum dependency attributes technique

The process of selecting a clustering attribute of MDA technique [19] is briefly described below:

Given m attributes in the information system $S = (U, A, V, f)$, Max-Dependency (MD) of attribute a_i, for $a_i \in A$ with respect to all attributes in S is defined as

$$\text{MD}(a_i) = \max\{k_{a_1}(a_i), \cdots, k_{a_j}(a_i), \cdots, k_{a_m}(a_i)\} \quad (3)$$

where $a_i \neq a_j$, $1 \leqslant i, j \leqslant m$

After obtaining the $|m|$ value of $\text{MD}(a_i)$, $i = 1, 2, \cdots, m$, MDA technique selects the attribute with the maximum value of MD as clustering attribute, i.e.

$$\text{MDA} = \max\{\text{MD}(a_1), \cdots, \text{MD}(a_i), \cdots, \text{MD}(a_m)\} \quad (4)$$

Based on the dependency of attributes in the rough set theory in the information system, MDA algorithm is given in Figure 1:

Example 3. From Example 2, we have the following dependencies among all attributes.

Like MMR technique, if the highest dependence degree of an attribute is the same with other attributes, then it is recommended to look at the next highest MDA inside the attributes until the tie is broken. From Table 3, it is clear that the attribute area is selected as a clustering attribute.

In reviewing MDA algorithm, it still is time-consuming in calculating the dependency degree of all attributes. In order to improve this situation, we propose a novel soft set approach in selecting a clustering attribute.

4. Proposed soft set-based technique

In this section, we introduce a Novel Soft Set Approach (NSSA) in selecting a clustering attribute. First, we define a soft set model on the equivalence classes for the information system, which can be easily applied to obtaining approximation of the rough set. Furthermore, we use it to select a clustering attribute and a heuristic algorithm is presented.

4.1. A soft set model on equivalence classes

There are many applications on the information system including computation related to equivalence classes or attributes. The computation may be intersection, union of the sets of equivalence classes or dependency degree among attributes. It is inconvenient to execute these operations directly on the information system; therefore, based on the basic definition of soft set described in Definition 5, we construct a soft set model over equivalence classes to facilitate the computation related to equivalence classes and attributes. It is defined as follows:

Definition 3. Let $S = (U, A, V, f)$ in the information system and U/A denotes the set of all equivalence classes in the partitions U/a_i, where $a_i \in A$. Let $U' = U/A$ be the initial universe of objects, $E = U/A$

be the set of parameters. $P(U')$ denotes the power set of U', and defines mapping F: $E \to P(U')$. We call the pair (F, E), a soft set model over equivalence classes.

From this definition, for any equivalence class $e \in E$, $F(e) \subseteq U'$ is the set of equivalence classes which have some certain relations with e. By defining different mapping F, we can construct different soft sets to meet various requirements. Table 4 shows the tabular representation of the soft set over equivalence classes.

Let $U' = E$, $e_i = x_i$, $i = 1, 2, \cdots, m$, where $m = |U/A| = |U'|$. Thus, $F(e_i)(x_i)$ is equal to 0 or 1. Based on the proposed soft set model, we construct two soft sets to compute lower and upper approximation sets of an equivalence class or attribute with respect to other attributes in rough set as follows.

Definition 4. Let $P(U')$ be the power set of a universe U'. We define two mappings F_1, F_2: $E \to P(U')$, for

(a) $F_1(e) = \{A | A \in U' \text{ and } A \subset e\}$.
(b) $F_2(e) = \{A | A \in U' \text{ and } A \cap e \neq \phi\}$.

For any equivalence class $e \in E$, $F_1(e) \subseteq U'$ is the set of equivalence classes which are subsets of e. Meanwhile, $F_2(e) \subseteq U'$ is the set of equivalence classes which have intersection with e. Having these two mappings, we can construct two soft sets (F_1, E) and (F_2, E). An illustrative example of the two soft sets is given in Example 4.

Example 4. We consider the information system shown in Table 2. From partitions in (2) and Definition 3, we have the following collection:

$$U/A = \{\{1,2,5\},\{3,4\},\{1,2\},\{3,5\},\{4\},\{1\},\{2,3,4,5\}\}. \quad (5)$$

We firstly construct the soft set (F_1, E). Since $U' = E = U/A$, from Definition 4 and a collection in (5), we can obtain the following value of mapping:

$$F_1(\{1,2,5\}) = \{\{1,2,5\},\{1,2\},\{1\}\},$$
$$F_1(\{3,4\}) = \{\{3,4\},\{4\}\},$$
$$F_1(\{1,2\}) = \{\{1,2\},\{1\}\},$$
$$F_1(\{3,5\}) = \{\{3,5\}\}$$
$$F_1(\{4\}) = \{\{4\}\},$$
$$F_1(\{1\}) = \{\{1\}\},$$
$$F_1(\{2,3,4,5\}) = \{\{3,4\},\{3,5\},\{4\},\{2,3,4,5\}\}.$$

The tabular representation of the soft set (F_1, E) is shown in Table 5.

Similarly, we can construct the soft set (F_2, E). Since $U' = E = U/A$, from Definition 4 and a collection in Eq. (5), we can obtain the following value of mapping

$$F_2(\{1,2,5\}) = \{\{1,2,5\},\{1,2\},\{1\},\{3,5\},\{2,3,4,5\}\},$$
$$F_2(\{3,4\}) = \{\{3,4\},\{4\},\{3,5\},\{2,3,4,5\}\},$$
$$F_2(\{1,2\}) = \{\{1,2\},\{1\},\{1,2,5\},\{2,3,4,5\}\},$$
$$F_2(\{3,5\}) = \{\{3,4\},\{3,5\},\{1,2,5\},\{2,3,4,5\}\}$$
$$F_2(\{4\}) = \{\{4\},\{3,4\},\{2,3,4,5\}\},$$
$$F_2(\{1\}) = \{\{1,2\},\{1\},\{1,2,5\}\},$$
$$F_2(\{2,3,4,5\}) = \{\{3,4\},\{1,2\},\{1,2,5\},\{3,5\},\{4\},\{2,3,4,5\}\}.$$

Table 3
Dependency degree of attributes in Table 2.

Attribute w.r.t	Shape	Color	Area	MDA
Shape	–	0.6	0.2	0.6
				0.2
Color	0	–	0.2	
Area	0.4	0.6	–	0.6
				0.4

Table 4
Tabular representation of the soft set over equivalence classes.

U'/E	e_1	e_2	e_3	...	e_m
x_1	$F(e_1)(x_1)$	$F(e_2)(x_1)$	$F(e_3)(x_1)$	...	$F(e_m)(x_1)$
x_2	$F(e_1)(x_2)$	$F(e_2)(x_2)$	$F(e_3)(x_2)$	...	$F(e_m)(x_2)$
x_3	$F(e_1)(x_3)$	$F(e_2)(x_3)$	$F(e_3)(x_3)$	...	$F(e_m)(x_3)$
...	...	...	...	...	...
x_m	$F(e_1)(x_m)$	$F(e_2)(x_m)$	$F(e_3)(x_m)$	...	$F(e_m)(x_m)$

Table 5
Tabular representation of the soft set (F_1, E).

U'/E	e_1{1, 2, 5}	e_2{3, 4}	e_3{1, 2}	e_4{3, 5}	e_5{4}	e_6{1}	e_7{2, 3, 4, 5}
x_1{1, 2, 5}	1	0	0	0	0	0	0
x_2{3, 4}	0	1	0	0	0	0	1
x_3{1, 2}	1	0	1	0	0	0	0
x_4{3, 5}	0	0	0	1	0	0	1
x_5{4}	0	1	0	0	1	0	1
x_6{1}	1	0	1	0	0	1	0
x_7{2, 3, 4, 5}	0	0	0	0	0	0	1

Table 6
The tabular representation of the soft set (F_2, E).

U'/E	e_1{1, 2, 5}	e_2{3, 4}	e_3{1, 2}	e_4{3, 5}	e_5{4}	e_6{1}	e_7{2, 3, 4, 5}
x_1{1, 2, 5}	1	0	1	1	0	1	1
x_2{3, 4}	0	1	0	1	1	0	1
x_3{1, 2}	1	0	1	0	0	1	1
x_4{3, 5}	1	1	0	1	0	0	1
x_5{4}	0	1	0	0	1	0	1
x_6{1}	1	0	1	0	0	1	0
x_7{2, 3, 4, 5}	1	1	1	1	1	0	1

Table 7
Degree of dependency of all attributes in Table 2.

Attribute w.r.t	Shape	Color	Area
Shape	–	0.6	0.2
Color	0	–	0.2
Area	0.4	0.6	–

The tabular representation of the soft set (F_2, E) is shown in Table 6.

Based on the soft sets (F_1, E) and (F_2, E), we build up a computational model for lower and upper approximation sets:

The *lower approximation of parameter* e_j with respect to attribute a_i is defined as

$$\underline{a_i}(e_j) = \{x | x \in x_k \text{ and } F_1(e_j)(x_k) = 1, k = D+1, \cdots, D+|U/a_i|\} \quad (6)$$

where D refers to the total number of equivalence classes in the partitions U/a_l, where $l = 1, 2, \cdots, i-1$, namely $D = \sum_{l=1}^{i-1}|U/a_l|$, $x_k \in U'$ is one of the equivalence classes induced by U/a_i, and $F_1(e_j)(x_k) \in \{0,1\}$ is the entry of the tabular representation of the soft set (F_1, E).

The *cardinality* of $\underline{a_i}(e_j)$ can be calculated as

$$|\underline{a_i}(e_j)| = \sum_{k=D+1}^{D+|U/a_i|} F_1(e_j)(x_k) \times |x_k|. \quad (7)$$

The *lower approximation of attribute* a_j with respect to attribute a_i is defined as

$$\underline{a_i}(a_j) = \{x | x \in \underline{a_i}(e_k), k = D+1, \cdots, D+|U/a_j|\}, \quad (8)$$

Algorithm 1: MDA

Input: Data set without a clustering attribute
Begin
Step 1: Compute the equivalence classes using the indiscernible relation on each attribute;
Step 2: Determine the dependency degree of attribute a_i with respect to all a_j, where $i \neq j$;
Step 3: Select the maximum dependency degree of each attribute;
Step 4: Select a clustering attribute based on the maximum degree of the dependency attributes.
Output: A clustering attribute
End

Fig. 1. MDA algorithm.

Algorithm 2: A novel soft set (NSS) approach in selecting clustering attribute

Input: Data set without a clustering attribute
Begin
Step 1: Compute the equivalence classes using the indiscernible relation on each attribute;
Step 2: Construct soft set (F_1, E);
Step 3: Construct the tabular representation of the soft set (F_1, E);
Step 4: Compute the cardinality of lower approximation of an attribute with respect to other attributes in terms of Equation (8);
Step 5: Compute the dependency of an attribute with respect to other attributes in terms of Equation (14);
Step 6: Select the attribute with the highest dependency as the clustering attribute.
Output: A clustering attribute
End

Fig. 2. The NSS algorithm.

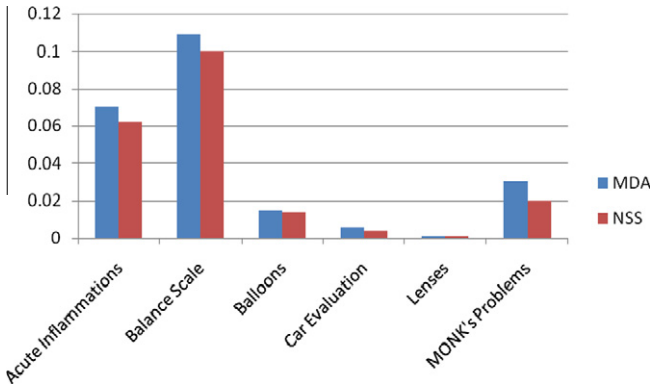

Fig. 3. The executing time (in seconds) of MDA and NSS.

where $D = \sum_{l=1}^{j-1}|U/a_l|$.

The *cardinality* of $\underline{a_i}(a_j)$ can be calculated as

$$|\underline{a_i}(a_j)| = \sum_{k=D+1}^{D+|U/a_j|} |\underline{a_i}(e_k)| \tag{9}$$

Similarly, the *upper approximation of parameter* e_j with respect to attribute a_i is defined as

$$\overline{a_i}(e_j) = \{x | x \in x_k \text{ and } F_2(e_j)(x_k) = 1, k = D+1, \cdots, D+|U/a_i|\} \tag{10}$$

where D has the same meaning as in Eq. (6), namely $D = \sum_{l=1}^{i-1}|U/a_l|$, $F_2(e_j)(x_k) \in \{0,1\}$ is the entry of the tabular representation of the soft set (F_2, E).

The *cardinality* of $\overline{a_i}(e_j)$ can be calculated as

$$|\overline{a_i}(e_j)| = \sum_{k=D+1}^{D+|U/a_i|} F_2(e_j)(x_k) \times |x_k|. \tag{11}$$

The *upper approximation of attribute* a_j with respect to attribute a_i is defined as

$$\overline{a_i}(a_j) = \{x | x \in \overline{a_i}(e_k), k = D+1, \cdots, D+|U/a_j|\}, \tag{12}$$

where $D = \sum_{l=1}^{j-1}|U/a_l|$.

The *cardinality* of $\overline{a_i}(a_j)$ can be calculated as

$$|\overline{a_i}(a_j)| = \sum_{k=D+1}^{D+|U/a_j|} |\overline{a_i}(e_k)| \tag{13}$$

With the computational model, it is convenient to compute the lower and upper approximations of equivalence class or attribute with respect to other attributes.

Let us reconsider Example 2 and suppose we are required to compute the cardinality of lower approximation of equivalence class $e_1 = \{1, 2, 5\}$ with respect to attribute 'Color', according to Eq. (7), we have

$$|\text{Color}(e_1)| = \sum_{k=3}^{5} F(e_1)(x_k) \times |x_k| = (1 \times 2) + (0 \times 2) + (0 \times 1) = 2.$$

Furthermore, if we are asked to compute the lower approximation of attribute 'Shape' with respect to attribute 'Color', we have

$$\underline{\text{Color}}(\text{Shape}) = \{x | x \in \underline{\text{Color}}(e_1), \underline{\text{Color}}(e_2)\} = \{1, 2, 4\}.$$

4.2. The NSS algorithm

From Eq. (1), it can be seen that only lower approximation is required in MDA technique, and we can further simplify Eq. (1) as

$$k_{a_j}(a_i) = \frac{|\underline{a_j}(a_i)|}{|U|} \tag{14}$$

We propose the algorithm based on soft set as given in Figure 2.

Let us reconsider Example 2 The first three steps have been shown in Example 2. We will start with the fourth step, namely, compute the cardinality of lower approximation in terms of Eq. (9). We can obtain

$$|\underline{\text{Color}}(\text{Shape})| = |\underline{\text{Color}}(e_1)| + |\underline{\text{Color}}(e_2)| = 2 + 1 = 3$$
$$|\underline{\text{Area}}(\text{Shape})| = |\underline{\text{Area}}(e_1)| + |\underline{\text{Area}}(e_2)| = 1 + 0 = 1$$
$$|\underline{\text{Shape}}(\text{Color})| = |\underline{\text{Shape}}(e_3)| + |\underline{\text{Shape}}(e_4)| + |\underline{\text{Shape}}(e_5)| = 0 + 0 + 0 = 0$$
$$|\underline{\text{Area}}(\text{Color})| = |\underline{\text{Area}}(e_3)| + |\underline{\text{Area}}(e_4)| + |\underline{\text{Area}}(e_5)| = 1 + 0 + 0 = 1$$
$$|\underline{\text{Shape}}(\text{Area})| = |\underline{\text{Shape}}(e_6)| + |\underline{\text{Shape}}(e_7)| = 0 + 2 = 2$$
$$|\underline{\text{Color}}(a_3)| = |\underline{\text{Color}}(e_6)| + |\underline{\text{Color}}(e_7)| = 0 + 3 = 3$$

Next, we compute the dependency degree of an attribute with respect to other attributes in terms of Eq. (14). The results are summarized in Table 7.

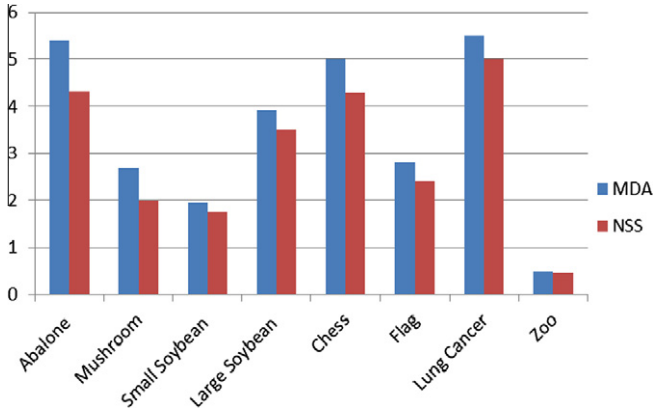

Fig. 4. The executing time (in seconds) of MDA and NSS.

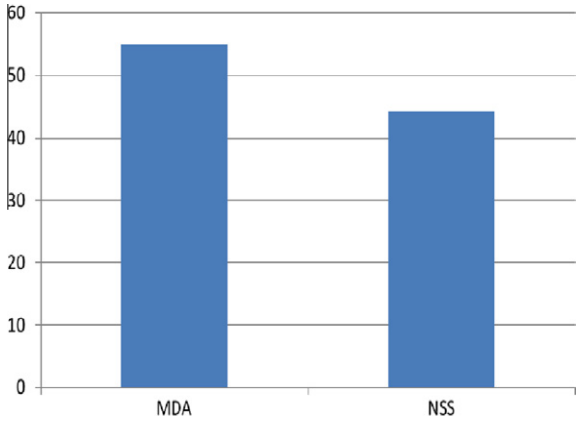

Fig. 5. The executing time (in seconds) of MDA and NSS on Statlog dataset.

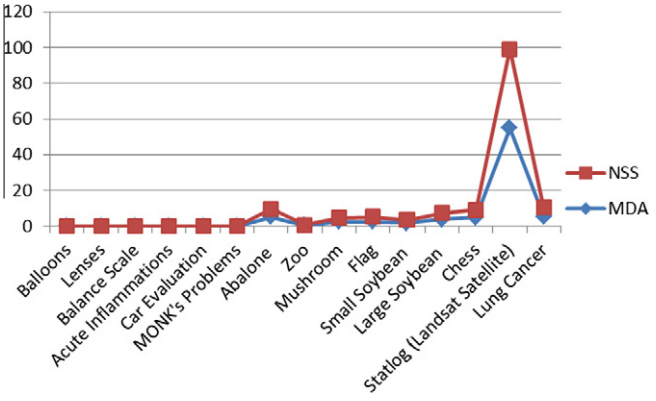

Fig. 7. The scalability of MDA and NSS to the number of attributes.

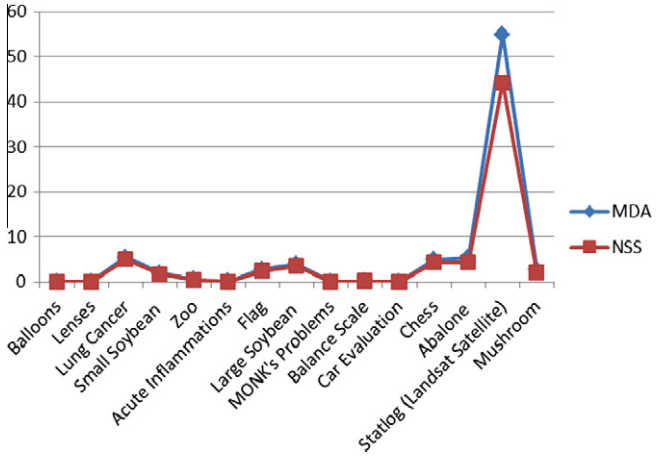

Fig. 6. The scalability of MDA and NSS to the number of instances.

Referring toTable 7, the attribute area will be selected as a clustering attribute.

5. Experimental results

In order to test the proposed NSS algorithm and compare it with MDA [19] in terms of executing time, we use fifteen datasets obtained from the benchmark UCI Machine Learning Repository [34]. The datasets are described in Table 8.

The MDA and NSS are implemented in Matlab version 7. They are sequentially executed on a PC with a processor Intel Core 2 Duo 2.0 GHz. The main memory is 2 GB and the operating system is Widows XP Professional. The results on experiments done on the datasets are given in Fig. 3 for executing time results less than 0.4 sec and Fig. 4 for executing time results more than 0.4 s and less than 10 s. Meanwhile Fig. 5 presents executing time results on Statlog dataset where it achieves more than 40 s.

Table 9 summarizes the comparison results in terms of executing time between MDA and NSS. It is clearly shown that NSS significantly improves MDA and on the average, NSS significantly outperforms MDA up to 14.84%.

Further, we present two types of scalability of MDA and NNS algorithms on the datasets above. Fig. 6 shows the executing time of using both algorithms to the number of instances (objects) in the datasets. Meanwhile, Fig. 7 shows the executing time through the datasets as number of attributes increasingly. It can be observed from Figs. 6 and 7 that the executing time of both algorithms tends to increase linearly as the number of instances and attributes are increased, respectively.

From Figs. 6 and 7, the value varies at several numbers of instances and attributes, respectively. However, the scalability is at an acceptable level on the whole. Thus, both MDA and NSS have a good scalability, they can be applied on small as well as large categorical data sets.

Table 8
Benchmark datasets.

No.	Data sets	Number of instances	Number of attributes
1	Abalone	4177	8
2	Acute inflammations	120	6
3	Balance scale	625	4
4	Balloons	16	4
5	Car evaluation	1728	6
6	Chess	3196	36
7	Flag	194	30
8	Lenses	24	4
9	Lung cancer	32	56
10	MONK's problems	432	7
11	Mushroom	8124	22
12	Small soybean	47	35
13	Large soybean	307	35
14	Statlog (landsat satellite)	6435	36
15	Zoo	101	17

Table 9
Summary of comparison between MDA and NSS.

No.	Data sets	Executing time (s)		Improvement (%)
		MDA	NSS	
1	Abalone	5.400	4.310	20.19
2	Acute inflammations	0.070	0.062	11.43
3	Balance scale	0.109	0.100	8.26
4	Balloons	0.015	0.014	6.67
5	Car evaluation	0.006	0.004	33.33
6	Chess	5.000	4.300	14.00
7	Flag	2.800	2.400	14.29
8	Lenses	0.001	0.001	0.00
9	Lung cancer	5.500	5.000	9.09
10	MONK's problems	0.030	0.020	33.33
11	Mushroom	2.700	2.000	25.93
12	Small soybean	1.950	1.750	10.26
13	Large soybean	3.900	3.500	10.26
14	Statlog (landsat satellite)	54.939	44.200	19.55
15	Zoo	0.500	0.470	6.00
Average improvement				14.84

6. Conclusion

In this paper, we have presented a soft set model on equivalence classes in the information system. Based on the proposed model, we design two soft sets in order to obtain approximation of the rough set. Furthermore, a Novel Soft Set (NSS) approach in selecting a clustering attribute is recommended. Experiment results on 15 UCI benchmark datasets signaled that the proposed approach provides a faster decision in selecting a clustering attribute, compared with rough set-based and Maximum Dependency Attributes (MDAs) techniques, which is up to 14.845%. Furthermore, MDA and NSS have a good scalability i.e. the executing time of both algorithms tends to increase linearly as the number of instances and attributes are increased, respectively.

In future, we will evaluate the proposed approach for categorical data clustering and compare it with the rough set-based clustering approach. This will show that the proposed approach can be used to solve practical problems in categorical data clustering.

Acknowledgment

This work was supported by Postgraduate Research Grant Scheme under the Grant No. GRS100323, Universiti Malaysia Pahang, Malaysia. The authors would like to thanks Dr. Andrew Tse for carefully proof reading this manuscript.

References

[1] M. Halkidi, Y. Batistakis, M. Vazirgiannis, On clustering validation techniques, Journal of Intelligent Information Systems 17 (2–3) (2001) 107–145.
[2] Z. Pawlak, Rough sets, International Journal of Computer and Information Science 11 (1982) 341–356.
[3] Z. Pawlak, Rough Sets: A Theoretical Aspect of Reasoning About Data, Kluwer Academic Publisher, 1991.
[4] Z. Pawlak, A. Skowron, Rudiments of rough sets, Information Sciences 177 (1) (2007) 3–27.
[5] L. An, L. Tong, Rough approximations based on intersection of indiscernibility, similarity and outranking relations, Knowledge-Based Systems 23 (6) (2010) 555–562.
[6] R. Yan, J. Zheng, J. Liu, Y. Zhai, Research on the model of rough set over dual-universes, Knowledge-Based Systems 23 (8) (2010) 817–822.
[7] L. Wei, J.-J. Qi, Relation between concept lattice reduction and rough set reduction, Knowledge-Based Systems 23 (8) (2010) 934–938.
[8] Y. Xu, L. Wang, R. Zhang, A dynamic attribute reduction algorithm based on 0–1 integer programming, Knowledge-Based Systems 24 (8) (2011) 1341–1347.
[9] L. Feng, T. Li, D. Ruan, S. Gou, A vague–rough set approach for uncertain knowledge acquisition, Knowledge-Based Systems 24 (6) (2011) 837–843.
[10] P. Yang, Q. Zhu, Finding key attribute subset in dataset for outlier detection, Knowledge-Based Systems 24 (2) (2011) 269–274.
[11] Q. He, C. Wu, D. Chen, S. Zhao, Fuzzy rough set based attribute reduction for information systems with fuzzy decisions, Knowledge-Based Systems 24 (5) (2011) 689–696.
[12] I.T.R. Yanto, T. Herawan, M.M. Deris, Data clustering using variable precision rough set, Intelligent Data Analysis 15 (4) (2011) 465–482.
[13] W. Xu, Y. Li, X. Liao, Approaches to attribute reductions based on rough set and matrix computation in inconsistent ordered information systems, Knowledge-Based Systems 27 (2012) 78–91.
[14] I.T.R. Yanto, P. Vitasari, T. Herawan, M.M. Deris, Applying variable precision rough set model for clustering student suffering study's anxiety, Expert System with Applications 39 (1) (2012) 452–459.
[15] D. Chen, D. Cui, C. Wang, Z. Wang, A rough set-based hierarchical clustering algorithm for categorical data, International Journal of Information Technology 12 (3) (2006) 149–159.
[16] L.J. Mazlack, A. He, Y. Zhu, S. Coppock, A rough set approach in choosing clustering attributes, in: The Proceedings of the ISCA 13th, International Conference CAINE-2000, 2000, pp. 1–6.
[17] D. Parmar, T. Wu, J. Blackhurst, MMR: an algorithm for clustering categorical data using rough set theory, Data and Knowledge Engineering 63 (2007) 879–893.
[18] Y.Y. Yao, Information granulation and rough set approximation, International Journal of Intelligent Systems 16 (1) (2001) 87–104.
[19] T. Herawan, M.M. Deris, J.H. Abawajy, A rough set approach for selecting clustering attribute, Knowledge-Based Systems 23 (2010) 220–231.
[20] D. Molodtsov, Soft set theory_first results, Computers and Mathematics with Applications 37 (1999) 19–31.
[21] L.A. Zadeh, Fuzzy sets, Information and Control 8 (1965) 338–353.
[22] D. Molodtsov, The Theory of Soft Sets, URSS Publishers, Moscow, 2004 (in Russian).
[23] F. Feng, C. Li, B. Davvaz, M.I. Ali, Soft sets combined with fuzzy sets and rough sets: a tentative approach, in: Soft Computing – A Fusion of Foundations, Methodologies and Applications, Springer Verlag, 2009, pp. 899–911.
[24] F. Feng, X.Y. Liu, V. Leoreanu-Fotea, Y.B. Jun, Soft sets and soft rough sets, Information Science 181 (2011) 1125–1137.
[25] T. Herawan, M.M. Deris, A direct proof of every rough set is a soft set, in: Proceeding of the Third Asia International Conference on Modeling and Simulation, Bali, Indonesia, AMS'09, IEEE Press, 2009, pp. 119–124.
[26] P.K. Maji, A.R. Roy, An application of soft sets in a decision making problem, Computers and Mathematics with Applications 44 (2002) 1077–1083.
[27] X. Ma, S. Norrozila, H. Qin, T. Herawan, J.M. Zain, A new efficient normal parameter reduction algorithm of soft sets, Computers and Mathematics with Applications 62 (2011) 588–598.
[28] F. Feng, Y.B. Jun, X.Y. Liu, L.F. Li, An adjustable approach to fuzzy soft set based decision making, Journal of Computational and Applied Mathematics 234 (2010) 10–20.
[29] F. Feng, Y.M. Li, N. Cagman, Generalized uni-int decision making schemes based on choice value soft sets, European Journal of Operational Research 220 (1) (2012) 162–170.
[30] Y. Zou, Z. Xiao, Data analysis approaches of soft sets under incomplete information, Knowledge Based System 21 (8) (2008) 941–945.
[31] H. Qin, X. Ma, T. Herawan, J.M. Zain, Data filling approach of soft sets under incomplete information, in: N.T. Nguyen, C.G. Kim, A. Janiak, (Eds.), ACIIDS 2011, vol. 6592, Lecture Notes in Computer Science, 2011, pp. 302–311.
[32] Z. Xiao, K. Gong, Y. Zou, A combined forecasting approach based on fuzzy soft sets, Journal of Computational and Applied Mathematics 228 (1) (2009) 326–333.
[33] T. Herawan, M. Mat Deris, A soft set approach for association rules mining, Knowledge Based Systems 24 (1) (2011) 186–195.
[34] UCI Repository of Machine Learning Databases, Retrieved from <http://www.ics.uci.edu/~mlearn/MLRRepository.html>.