A Soft Set Model on Information System and Its Application in Clustering Attribute Selection

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

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

Abstract. In this paper, we define a soft set model on the set of equivalence classes in an information system, which can be easily applied to obtaining approximation sets of rough set. Furthermore, we use it to select clustering attribute for categorical data clustering and a heuristic algorithm is presented. Experiment results on UCI benchmark data sets show that the proposed approach provides faster decision for selecting a clustering attribute as compared with maximum dependency attributes (MDA) approach.

Keywords: Soft set, Rough set, Information system, Clustering attribute.

1 Introduction

In 1999, Molodtsov [1] proposed soft set theory as a new mathematical tool for dealing with vagueness and uncertainties. Where soft set theory is different from traditional tools for dealing with uncertainties, such as the theory of probability, the theory of fuzzy sets, is that it is free from the inadequacy of the parametrization tools of those theories. 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. [4] investigated the problem of combining soft sets with fuzzy sets and rough 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 give a direct proof in [5] that every rough set is a soft set.

Rough set theory [2], 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. The original goal of the rough set theory is induction of approximations of concepts. The idea consists of approximation of a subset by a pair of two precise concepts called the lower approximation and upper approximation. Intuitively, the lower approximation of a set consists of all elements that surely belong to the set, whereas the upper approximation of the set constitutes of

J.M. Zain et al. (Eds.): ICSECS 2011, Part II, CCIS 180, pp. 16–27, 2011.

all elements that possibly belong to the set. The difference of the upper approximation and the lower approximation is a boundary region. It consists of all elements that cannot be classified uniquely to the set or its complement, by employing available knowledge.

In this paper, we propose a soft set model on information system. The soft set model is constructed over the set of equivalence class instead of the set of single object. Then we apply the soft set model to obtaining approximation sets of rough set. Furthermore, we use it to select clustering attribute for categorical data cluster and a heuristic algorithm is presented. Experiment results on UCI benchmark data sets show that the proposed approach provides faster decision for selecting a clustering attribute as compared with maximum dependency attributes (MDA) approach.

The rest of this paper is organized as follows. The following section briefly reviews some basic definitions in rough set theory and soft sets. Section 3 describes the construction of soft set model on information system. Section 4 shows the application of the proposed model in Selecting Clustering Attributes. Section 5 makes comparison between MDA approach and the proposed technique. Finally, conclusions are given in Section 6.

2 Preliminaries

The notion of information system provides a convenient tool for the representation of objects in terms of their attribute values. An information system as in [3] is a 4-tuple $S = (U, A, V, f)$, where U is a non-empty finite set of objects, 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 a total function such that $f(u,a) \in V_a$, for every $(u,a) \in U \times A$, called information function. Next, we review some basic definitions with regard to rough set and soft set.

Definition 1. *Let $S = (U, A, V, f)$ be an information system and let B be any subset of A. Two elements $x, y \in U$ 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 $a \in B$.*

An indiscernibility relation induced by the set of attribute B, denoted by IND (B), is an equivalence relation. It is well-known that, an equivalence relation induces 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 containing $x \in U$, denoted by $[x]_B$.

Definition 2. *Let $S = (U, A, V, f)$ be an information system, let B be any subset of A 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*

$$\underline{B}(X) = \{x \in U \mid [x]_B \subseteq X\} \text{ and } \overline{B}(X) = \{x \in U \mid [x]_B \cap X \neq \phi\} \tag{1}$$

The accuracy of approximation of any subset $X \subseteq U$ with respect to $B \subseteq A$, denoted by $\alpha_B(X)$ is measured by

$$\alpha_B(X) = \frac{|\underline{B}(X)|}{|\overline{B}(X)|} \tag{2}$$

Throughout the paper, |*X*| denotes the cardinality of *X*.

Definition 3. *Let S = (U, A, V, f) be an information system and let G and H be any subsets of A. G depends on H in a degree k, is denoted by* $H \Rightarrow_k G$. *The degree k is defined by*

$$k = \frac{\sum_{X \in U/G} |\underline{H}(X)|}{|U|} \tag{3}$$

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

Definition 4. *A pair (F, E) is called a soft set over U, where F is a mapping given by*

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

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

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

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

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

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

3 A Soft Set Model on Equivalence Classes

There are many applications on 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 as in Definition 4, we construct a soft set model over equivalence classes to facilitate the computation related to equivalence classes and attributes. The soft set model is defined as follows.

Definition 5. *Given $S = (U, A, V, f)$ be an information system, let U/A denotes the set of all equivalence classes in the partitions U/a_i ($a_i \in A$ and $i = 1,2,\ldots,|A|$). 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 define mapping $F: E \rightarrow P(U')$, 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 2 shows the tabular representation of the soft set over equivalence classes.

Table 2. Tabular representation of the soft set over equivalence classes

U'	e_1	e_2	e_i	…	e_m
x_1	$F(e_1)(x_1)$	$F(e_2)(x_1)$	$F(e_i)(x_1)$	…	$F(e_m)(x_1)$
x_2	$F(e_1)(x_2)$	$F(e_2)(x_2)$	$F(e_i)(x_2)$	…	$F(e_m)(x_2)$
x_i	$F(e_1)(x_i)$	$F(e_2)(x_i)$	$F(e_i)(x_i)$	…	$F(e_m)(x_i)$
…	…	…	…	…	…
x_m	$F(e_1)(x_m)$	$F(e_2)(x_m)$	$F(e_i)(x_m)$	…	$F(e_m)(x_m)$

Because $U'=E$, hence in Table 2 $e_i = x_i$ (i=1, …, m), where $m = |U/A| = |U'|$. $F(e_i)(x_i)$ equal 0 or 1.

Based on the proposed soft set model, in detail, 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.

We define two mappings F_1, F_2: $E \to P(U')$ as follows,

(1) F_1 : *Subsetof* (E)
(2) F_2 : *HasIntersectionWith* (E)

For any equivalence class $e \in E$, $F_1(e) \subseteq U'$ is the set of equivalence classes which are subsets of e, $F_2(e) \subseteq U'$ is the set of equivalence classes which have intersection with e. Having the 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 2.

Example 2. We consider the information system as shown in Table 3, where $U = \{x_1 =1, x_2 =2, x_3 =3, x_4 =4, x_5 =5\}$, $A = \{a_1$ =Shape, a_2 =Color, a_3 =Area$\}$.

Table 3. An information system of objects' appearance

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

From Table 3, there are three partitions of U induced by indiscernibility relation on each attribute, i.e.

$U/a_1 = \{\{1,2,5\},\{3,4\}\}$, $U/a_2 = \{\{1,2\}, \{3,5\},\{4\}\}$, $U/a_3 = \{\{1\},\{2,3,4,5\}\}$.

We firstly construct the soft set (F_1, E). We have

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

Consequently,
$U' = E = U/A$, and then we can obtain

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

The tabular representation of the soft set (F_1, E) is showed in Table 4.

Table 4. The tabular representation of the soft set (F_1, E)

U'	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

Similarly, we can construct the soft set (F_2, E). We have

U/A = {{1,2,5},{3,4}, {1,2}, {3,5},{4}, {1},{2,3,4,5}}.

U′ = *E* = *U/A*, and then we can obtain

F_2({1,2,5})=*HasIntersectionWith* ({1,2,5}) ={{1,2,5},{1,2}, {3,5},{1},{2,3,4,5}};
F_2 ({3,4}) = *HasIntersectionWith* ({3,4}) ={{3,4},{3,5},{4},{2,3,4,5}};
F_2 ({1,2}) = *HasIntersectionWith* ({1,2}) ={{1,2,5},{1,2},{1},{2,3,4,5}};
F_2 ({3,5}) = *HasIntersectionWith* ({3,5}) ={{1,2,5},{3,4},{3,5},{2,3,4,5}};
F_2 ({4}) = *HasIntersectionWith* ({4}) ={{3,4},{4},{2,3,4,5}};
F_2 ({1}) = *HasIntersectionWith* ({1}) ={{1,2,5},{1,2},{1}};
F_2 ({2,3,4,5}) = *Subsetof*({2,3,4,5}) ={{1,2,5}, {3,4},{1,2},{3,5},{4},{2,3,4,5}}.

The tabular representation of the soft set (F_2, E) is showed in Table 5.

Table 5. The tabular representation of the soft set (F_2, E)

U'	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

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

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

$$\underline{a_i}(e_j) = \{x | x \in x_k \quad and F_1(e_j)(x_k) = 1, k = D+1,\ldots, D+|U/a_i|\} \tag{4}$$

where D refers to the total number of equivalence classes in the partitions *U/a*$_l$(l=1,…,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*$_i$, $F_1(e_j)(x_k) \in \{0,1\}$ 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

$$\left|\underline{a_i}(e_j)\right| = \sum_{k=D+1}^{D+|U/a_i|} F_1(e_j)(x_k) \cdot |x_k| \tag{5}$$

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

$$\underline{a_i}(a_j) = \{x | x \in \underline{a_i}(e_k), k = D+1,\ldots,D+|U/a_j|\} \tag{6}$$

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

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

$$\left|\underline{a_i}(a_j)\right| = \sum_{k=D+1}^{D+|U/a_j|} \left|\underline{a_i}(e_k)\right| \tag{7}$$

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

$$\overline{a_i}(e_j) = \{x | x \in x_k \quad and F_2(e_j)(x_k) = 1, k = D+1,\ldots,D+|U/a_i|\} \tag{8}$$

where D has the same meaning as in Eq. (4), 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

$$\left|\overline{a_i}(e_j)\right| = \sum_{k=D+1}^{D+|U/a_i|} F_2(e_j)(x_k) \cdot |x_k| \tag{9}$$

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

$$\overline{a_i}(a_j) = \{x | x \in \overline{a_i}(e_k), k = D+1,\ldots,D+|U/a_j|\} \tag{10}$$

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

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

$$\left|\overline{a_i}(a_j)\right| = \sum_{k=D+1}^{D+|U/a_j|} \left|\overline{a_i}(e_k)\right| \tag{11}$$

With the computational model, it is convenient to compute the lower and upper approximation of equivalence class or attribute with respect to other attributes. Let us reconsider Example 2. Suppose we are required to compute the cardinality of lower approximation of equivalence class e_1 = {1, 2, 5} with respect to attribute a_2, according to Eq. (5), we have

$$\left|\underline{a_2}(e_1)\right| = \sum_{k=3}^{5} F(e_1)(x_k)\cdot\left|x_k\right| = 1\times 2 + 0\times 2 + 0\times 1 = 2 \,.$$

Furthermore, if we are required to compute the lower approximation of attribute a_1 with respect to attribute a_2, we have

$$\underline{a_2}(a_1) = \{x \mid x \in \underline{a_2}(e_1), \underline{a_2}(e_2)\} = \{1,2,4\}.$$

4 Application of the Proposed Model in Clustering Attribute Selection

Rough set theory based attribute selection clustering approaches for categorical data have attracted much attention in recent years. Mazlack et al. [7] proposed a technique using the average of the accuracy of approximation in the rough set theory called total roughness (TR). Parmar et al. [8] proposed a technique 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 an information system [9]. Herawan et al. [10] proposed a new technique called maximum dependency attributes (MDA) for selecting clustering attribute, which is based on rough set theory by taking into account the dependency of attributes of the database. Compared to TR and MMR, MDA technique provides better performance.

In this section, first, we briefly introduce MDA technique, and then propose an algorithm based on the soft set constructed in Section 5, finally the comparison tests between soft set based algorithm and MDA technique are implemented on two UCI benchmark data sets.

4.1 MDA Technique

In information system $S = (U, A, V, f)$, given any attribute a_i, a_j, $k_{a_j}(a_i)$ refer to a_i depends on a_j in degree k, is obtained by Eq. (4) as follows

$$k_{a_j}(a_i) = \frac{\sum_{X\in U/a_i}\left|\underline{a_j}(X)\right|}{|U|} \tag{12}$$

Next, given m attributes, *Max-Dependency* (MD) of attribute $a_i (a_i \in A)$ is defined as

$$MD(a_i) = Max(k_{a_1}(a_i),\ldots,k_{a_j}(a_i),\ldots,k_{a_m}(a_i)) \tag{13}$$

where $a_i \neq a_j$, $1 \leq i, j \leq m$.

After obtaining the m values of $MD(a_i)$, $i = 1,2,...,m$. MDA technique selects the attribute with the maximum value of MD as clustering attribute, i.e.

$$MDA = Max(MD(a_1),...,MD(a_i),...,MD(a_m)) \tag{14}$$

4.2 An Algorithm Based on Soft Set (F1, E)

From Eq. (12), it can be seen that only lower approximation is required in MDA technique, and we can further simplify Eq. (12) as

$$k_{a_j}(a_i) = \frac{\left|\underline{a_j}(a_i)\right|}{|U|} \tag{15}$$

We propose the algorithm based on soft set (F_1, E) as follows,

Algorithm 1

1. *Compute the equivalence classes using the indiscernibility relation on each attribute.*
2. *Construct soft set (F_1, E).*
3. *Construct the tabular representation of the soft set (F_1, E).*
4. *Compute the cardinality of lower approximation of an attribute with respect to other attributes in terms of Eq. (7).*
5. *Compute the dependency of an attribute with respect to other attributes in terms of Eq. (15).*
6. *Select the attribute with the highest dependency as the clustering attribute.*

Let us reconsider the Example 2 following Algorithm 1. The first three steps has been shown in Example 2, we start with the fourth step, namely compute the cardinality of lower approximation in terms of Eq. (7). We can obtain

$$\left|\underline{a_2}(a_1)\right| = \left|\underline{a_2}(e_1)\right| + \left|\underline{a_2}(e_2)\right| = 2 + 1 = 3$$

$$\left|\underline{a_3}(a_1)\right| = \left|\underline{a_3}(e_1)\right| + \left|\underline{a_3}(e_2)\right| = 1 + 0 = 1$$

$$\left|\underline{a_1}(a_2)\right| = \left|\underline{a_1}(e_3)\right| + \left|\underline{a_1}(e_4)\right| + \left|\underline{a_1}(e_5)\right| = 0 + 0 + 0 = 0$$

$$\left|\underline{a_3}(a_2)\right| = \left|\underline{a_3}(e_3)\right| + \left|\underline{a_3}(e_4)\right| + \left|\underline{a_3}(e_5)\right| = 1 + 0 + 0 = 1$$

$$\left|\underline{a_1}(a_3)\right| = \left|\underline{a_1}(e_6)\right| + \left|\underline{a_1}(e_7)\right| = 0 + 2 = 2$$

$$\left|\underline{a_2}(a_3)\right| = \left|\underline{a_2}(e_6)\right| + \left|\underline{a_2}(e_7)\right| = 0 + 3 = 3$$

Next, we compute the dependency degree of an attribute with respect to other attributes in terms of Eq. (15). The results are summarized in Table 6.

Table 6. The degree of dependency of all attributes in Table 3

w.r.t	a_1	a_2	a_3
a_1	-	0.6	0.2
a_2	0	-	0.2
a_3	0.4	0.6	-

Taking a look at Table 6, attribute a_3 (Area) will be selected as clustering attribute.

5 Experimental Results

In order to test Algorithm 1 and compare with MDA technique in [10], we use two datasets Soybean and Zoo obtained from the benchmark UCI Machine Learning Repository [6]. The two methods are implemented in C++ language. They are sequentially executed on a PC with a processor Intel Core 2 Duo 2.0GHz. The main memory is 2 GB and the operating system is Widows XP Professional SP3.

5.1 Soybean Data Set

The Soybean data set contains 47 instances on diseases in soybeans. Each instance is described by 35 categorical attributes and can be classified as one of the four diseases namely, Diaporthe Stem Canker, Charcoal Rot, Rhizoctonia Root Rot, and Phytophthora Rot. The data set is comprised 17 objects for Phytophthora Rot disease and 10 objects for each of the remaining diseases. Fig.1 illustrates the executing time of selecting the clustering attribute.

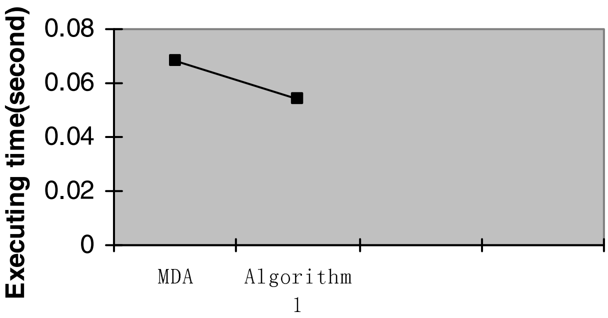

Fig. 1. The executing time of MDA and Algorithm 1on Soybean data set

5.2 Zoo Data Set

The Zoo data set contains 101 instances, where each instance represents information of an animal in terms of 16 categorical attributes. Each animal can be classified into seven classes namely, Mammal, Bird, Reptile, Fish, Amphibian, Insect, and Invertebrate. The data set is comprised 41 mammals, 20 birds, 5 reptiles, 13 fish, 4 amphibians, 8 insects and 10 invertebrates. Fig.2 illustrates the executing time of selecting the clustering attribute.

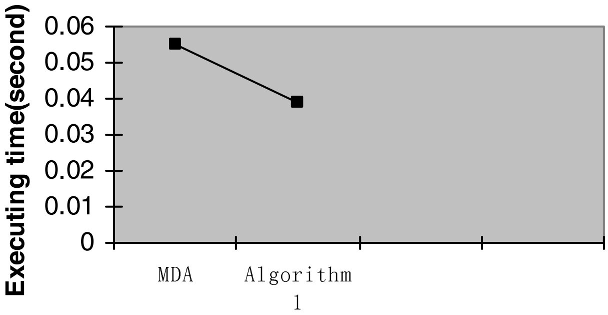

Fig. 2. The executing time of MDA and Algorithm 1on Zoo data set

From the above two experiments, it can be seen that Algorithm 1 improves the executing time of original MDA method.

6 Conclusions

In this paper, we present a soft set model on the set of equivalence classes in an information system. Based on the proposed model, in detail, we design two soft sets in order to obtain approximation sets of rough set. Furthermore, we make use of the proposed model to select clustering attribute for categorical data cluster and then a heuristic algorithm is presented. Experiment results on UCI benchmark data sets show that the proposed approach provides faster decision for selecting a clustering attribute as compared with maximum dependency attributes (MDA) approach.

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

References

1. Molodtsov, D.: Soft set theory_first results. Comput. Math. Appl. 37, 19–31 (1999)
2. Pawlak, Z.: Rough sets. International Journal Information Computer Science 11, 341–356 (1982)
3. Pawlak, Z., Skowron, A.: Rudiments of rough sets. Information Sciences: An International Journal 177(1), 3–27 (2007)
4. Feng, F., Li, C., Davvaz, B., Ali, M.I.: Soft sets combined with fuzzy sets and rough sets: a tentative approach. In: Soft Computing - A Fusion of Foundations, Methodologies and Applications, pp. 899–911. Springer, Heidelberg (2009)
5. Herawan, T., Deris, M.M.: A direct proof of every rough set is a soft set. In: Proceeding of the Third Asia International Conference on Modeling and Simulation, pp. 119–124 (2009)

6. UCI Repository of Machine Learning Databases, `http://www.ics.uci.edu/~mlearn/MLRRepository.html`
7. Mazlack, L.J., He, A., Zhu, Y., Coppock, S.: A rough set approach in choosing clustering attributes. In: Proceedings of the ISCA 13th International Conference (CAINE 2000), pp. 1–6 (2000)
8. Parmar, D., Wu, T., Blackhurst, J.: MMR: an algorithm for clustering categorical data using rough set theory. Data and Knowledge Engineering 63, 879–893 (2007)
9. Yao, Y.Y.: Information granulation and rough set approximation. International Journal of Intelligent Systems 16(1), 87–104 (2001)
10. Herawan, T., Deris, M.M., Abawajy, J.H.: A rough set approach for selecting clustering attribute. Knowledge-Based Systems 23, 220–231 (2010)