Discovering Attributes Dependency for Categorical Data Set Based on Soft Set Theory for Better Decision Making

Mohd Isa Awang\textsuperscript{*1}, Ahmad Nazari Mohd Rose\textsuperscript{1}, Fadhillah Ahmad\textsuperscript{1}, Mustafa Mat Deris\textsuperscript{2} and Tutut Herawan\textsuperscript{3}

\textsuperscript{1}Faculty of Informatics and Computing
Universiti Sultan Zainal Abidin, Tembila Campus, 22000 Besut, Terengganu Malaysia
\textsuperscript{2}Corresponding author

\textsuperscript{2}Faculty of Computer Science and Information Technology
Universiti Tun Hussein Onn Malaysia, 86400 Batu Pahat, Johor, Malaysia

\textsuperscript{3}Faculty of Computer Science and Information Technology
Universiti Malaya, 50603 Kuala Lumpur, Malaysia

Copyright © 2015 Mohd Isa Awang et al. This article is distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract

Attribute dependency concludes the association between attributes for better accurate decision making. However, the task involved in identifying the relation between categorical values in data set is a complex process. This main focus of this paper is to determine the attribute dependency in a real world application. The proposed method is based on the notion of mapping inclusion from the soft set theory. The categorical data is transformed to predicate and value set to discover the dependency among the attributes. The result shows that the attribute dependencies obtained are comparable to the rough set approach.

Keywords: Categorical data set; Attribute dependency; Soft set theory; Association rules; Decision making
1. Introduction

Attribute dependency is significant when uncovering the relationship among attributes in information systems for an improved decision making process. The rough set theory [1] has been extensively used for discovering attribute dependencies in information systems. It also has been applied in the many other domains; association rule mining [10,14,15,16,17,18], database management [21], reverse engineering, data clustering [5,6,7] and a few others. Any information system \( S=(U,A,V,f) \)[2], attribute \( D \) totally depends on attribute \( C \) if each value of \( D \) is associated exactly one value of \( C \), and this is denoted as \( C \Rightarrow D \). Otherwise, \( D \) is partially dependent on \( C \).

An attribute dependency defines that the value of an attribute is distinctively determined by the values of some other attributes.

One of the methods for discovering attribute dependencies is using rough set theory [1]. Formally, in information system \( S=(U,A,V,f) \)[2], attribute \( D \) is called totally depends on attribute \( C \), denoted \( C \Rightarrow D \), if each value of \( D \) is associated exactly one value of \( C \). Otherwise, \( D \) is depends partially on \( C \). The discovery of attribute dependencies using rough set theory has been received considerable interest (e.g., [3, 4, 5, 6, 7, 8, 9, 10]). Soft set theory [11], proposed by Molodtsov in 1999, is a new method for dealing with uncertain data. In recent years, research on soft set theory has been active, and great progress has been achieved including the works using fundamental of soft set theory, soft set theory in abstract algebra and soft set theory for data analysis, particularly in decision making.

The fact that every rough set is a soft set [12], has motivated us to study the applicability of soft set theory in discovering attributes dependencies in information systems. The suggested technique relies on the notion of a mapping inclusion in soft set theory. The obtained results are alike to the rough dependencies. Further, we present the application of such dependency for maximal association rules mining, decision making in a multi-valued domain and categorical data clustering.

The following content of this paper is ordered as follows. The fundamental concept of information systems and set approximations will be described in the Section 2. Section 3 explains the main concept of soft set theory. Section 4 depicts soft set approach for discovering attributes dependencies. Section 5 describes the three applications of soft attributes dependencies. The final section, Section 6 consists our conclusion.
2. Information Systems and Set Approximations

An information system \( S = (U, A, V, f) \) as in [2] is a 4-tuple (quadruple), where \( U \) and \( A \) is a non-empty finite set of objects and finite set of attributes respectively. While \( V_a \) from \( V = \bigcup_{a \in A} V_a \) is the domain (value set) of attribute \( a \), and \( f : U \times A \rightarrow V \) is a total function such that \( f(u, a) \in V_a \), for every \( (u, a) \in U \times A \), called information (knowledge) function. The outcome of classification from most application is already known, this \textit{a posteriori} knowledge is stated by one (or more) distinguished attribute called decision attribute; the process that is recognized as supervised learning. An information system of this kind is called a decision system.

A decision system is an information system of the form 

\[
D = (U, A \cup \{d\}, V, f),
\]

where \( d \in A \) is the decision attribute [2]. The elements of \( A \) are called condition attributes.

The indiscernibility relation as the initial point of rough set, is generated by information about objects of interest. Two objects in an information system are identified as indiscernible (or similar) if they have the same feature.

**Definition 1.** (See [2].) Two elements \( x, y \in U \) are said to be \( B \)-indiscernible if and only if \( f(x, a) = f(y, a) \), for every \( a \in B \).

Every subset of \( A \) will then induce unique indiscernibility relation. As an example, an indiscernibility relation induced by the set of attribute \( B \), symbolized by \( \text{IND}(B) \), will be known as equivalence relation. While \( U / B \) carries the meaning of a partition of \( U \) induced by \( \text{IND}(B) \). While the equivalence class in the partition \( U / B \) that contains \( x \in U \), is represented by \( [x]_B \). The lower and upper approximations of a set are characterized by the following definition:-

**Definition 2.** (See [2.]) \( B(X) \) denotes the \( B \)-lower approximation of \( X \), and \( B \)-upper approximations, denoted by \( \overline{B}(X) \) of \( X \), respectively, are defined by

\[
B(X) = \{ x \in U \mid [x]_B \subseteq X \} \quad \text{and} \quad \overline{B}(X) = \{ x \in U \mid [x]_B \cap X \neq \emptyset \}.
\]

The accuracy of roughness approximation of any subset \( X \subseteq U \) with respect to the set of attributes \( B \subseteq A \), denoted \( \alpha_B(X) \) is numerically measured by

\[
\alpha_B(X) = \frac{|B(X)|}{|\overline{B}(X)|},
\]

where \( |X| \) denotes the cardinality of \( X \). The higher of accuracy of approximation of any subset \( X \subseteq U \) is, the more precise (the less imprecise) of itself.

**Definition 3.** A rough approximation of a subset \( X \subseteq U \) with respect to \( B \) is defined as a pair of lower and upper approximations of \( X \), i.e. \( \langle B(X), \overline{B}(X) \rangle \).
3. Soft Set Theory

Throughout this section $U$ refers to an initial universe, $E$ is a set of parameters, $P(U)$ is the power set of $U$ and $A \subseteq E$.

**Definition 4.** (See [11].) A pair $(F, A)$ is called a soft set over $U$, where $F$ is a mapping given by $F : A \rightarrow P(U)$.

In other words, a soft set over $U$ is a parameterized family of subsets of the universe $U$. For $\varepsilon \in A$, $F(\varepsilon)$ may be considered as the set of $\varepsilon$-elements of the soft set $(F, A)$ or as the set of $\varepsilon$-approximate elements of the soft set. Clearly, a soft set is not a (crisp) set. For illustration, Molodtsov considered several examples in [11]. In this section, we will also present the relation between soft sets and rough sets, i.e., that every rough set is a soft set as stated by the following proposition.

**Proposition 5.** (see [12].) Every rough set can be considered as a soft set.

Let $\langle \underline{b}(X), \overline{b}(X) \rangle$ be the rough approximating a subset $X \subseteq U$. We define a mapping $\underline{b}, \overline{b} : P(U) \rightarrow P(U)$ as in Definition 3. Thus every rough set $\langle \underline{b}(X), \overline{b}(X) \rangle$ can be considered a pair of two soft sets $(F, U) = \langle \underline{b}(P(U)), \overline{b}(P(U)) \rangle$. □

We may see the structure of “standard” soft set usually classifies the objects into two classes (yes/1 or no/0). However, in the real application, a given parameter may not be restricted to only values of 0 and 1. That notion has been catered in multi-valued information system so-called multi-soft sets [13].

3.1 Multi-soft sets in information systems

The notion of a novel data structure called multi soft sets is illustrated by the delineation of $S = (U, A, V, f)$ as a multi-valued information system into $|A|$ number of Boolean-valued information systems $S^i = (U, a_i, V_{[0,1]}, f)$ [12]. Let $A = \{a_1, a_2, \ldots, a_{|A|}\}$, for every $a_i$ under $i$th-attribute consideration, $a_i \in A$ and $v \in V_{a_i}$, we define the map $a_i^v : U \rightarrow \{0,1\}$ such that $a_i^v(u) = 1$ if $f(u, a_i) = v$, otherwise $a_i^v(u) = 0$. Thus, we have

$$S = (U, A, V, f) = \left\{ \begin{array}{c}
S^1 = (U, a_1, V_{[0,1]}, f) \Leftrightarrow (F, a_1) \\
S^2 = (U, a_2, V_{[0,1]}, f) \Leftrightarrow (F, a_2) \\
\vdots \\
S^{|A|} = (U, a_{|A|}, V_{[0,1]}, f) \Leftrightarrow (F, a_{|A|})
\end{array} \right\}$$
The multi-pairs \((F,E) = [(F,a_1),(F,a_2),\ldots,(F,a_n)]\) is called a multi-soft sets over \(U\) representing \(S = (U,A,V,f)\). With this approach, we can discover attributes dependencies in a multi-valued information system using soft set theory. The proposed approach is presented in the following section.

4. Discovering Attributes Dependencies by Soft Set Approach

**Definition 6.** Let the representation of an information system \(S = (U,A,V,f)\) of \((F,E)\) to be a multi-soft sets over \(U\). While \(E_1,E_2 \subseteq E\) to be as the two different single attributes. The dependency degree \(k\) of \(E_2\) on \(E_1\) is defined by \(k = \left| F(E_1) : F(E_1) \subseteq F(E_2) \right|/|U|\).

Obviously, \(0 \leq k \leq 1\). Attribute \(D\) can be defined as to be totally dependent on the attribute \(C\) if \(k = 1\). Otherwise, \(D\) is depends partially on \(C\). The above definition is equivalent to the notion of rough attributes dependency in [2].

Soft set attribute dependency analysis is allowed to quantify numerically association rules and derive the best decision on the basis of information in data. However, there is no much literature on this topic. In this section, we apply such dependencies for maximal association rule mining under a Boolean-valued information system which is a soft set can be directly constructed and decision making under multi-valued domain using multi-soft sets approach.

5. Applications

5.1 Maximal association rules mining

Feldman et al. [14] had introduced the maximal association rule and it have been studied by many authors. While regular association rule [15] is based on the notion that in lots of records, there are item sets that appear frequently. And frequent maximal item sets which appears maximally in many records will define the rules for then maximal association. Bell et al. [11], Guan et al. [17] and Bi et al. [18] suggested the similar approach to determine the maximal association rules using rough set theory. Their recommended approach is focused on a partition on the set of all attributes in a transactional database so-called a taxonomy and categorization of such taxonomy. Let \(I = \{i_1,i_2,i_3,\ldots,i_n\}\) be a set of items and \(D = \{t_1,t_2,\ldots,t_m\}\) is transactional database over \(I\). A taxonomy \(T\) of \(I\) is a partition of \(I\) into disjoint sets \(T = \{T_1,T_2,\ldots,T_k\}\). Also, every element of \(T\) is defined as category. To those notions, it is clear that the pre-requisite of their approach is a transformation of a transactional database (bit-map table) into taxonomy-categorized table. In [14, 16, 17, 18], they are based on a widely used Boolean-valued information from Reuters-21578 [19] a labeled document collection, i.e. a benchmark for text categorization, as follows. Assume that the product corn have 10 articles which is related to the countries USA and Canada and 20 other publica-
tions regarding fish product and the countries USA, Canada and France. The information system is given in the following table.

Table 1. An information system from [9]

<table>
<thead>
<tr>
<th>U/A</th>
<th>US</th>
<th>Can</th>
<th>Fra</th>
<th>Corn</th>
<th>Fish</th>
</tr>
</thead>
<tbody>
<tr>
<td>$u_1$</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>$u_2$</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>$u_9$</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>$u_{10}$</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>$u_{11}$</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>$u_{12}$</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>$u_{29}$</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>$u_{30}$</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 2. The transformed table.

<table>
<thead>
<tr>
<th>U/A</th>
<th>Countries</th>
<th>Topics</th>
</tr>
</thead>
<tbody>
<tr>
<td>$u_1$</td>
<td>{US, Can}</td>
<td>{Corn}</td>
</tr>
<tr>
<td>$u_2$</td>
<td>{US, Can}</td>
<td>{Corn}</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>$u_9$</td>
<td>{US, Can}</td>
<td>{Corn}</td>
</tr>
<tr>
<td>$u_{10}$</td>
<td>{US, Can}</td>
<td>{Corn}</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>$u_{11}$</td>
<td>{US, Can, Fra}</td>
<td>{Fish}</td>
</tr>
<tr>
<td>$u_{12}$</td>
<td>{US, Can, Fra}</td>
<td>{Fish}</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>$u_{29}$</td>
<td>{US, Can, Fra}</td>
<td>{Fish}</td>
</tr>
<tr>
<td>$u_{30}$</td>
<td>{US, Can, Fra}</td>
<td>{Fish}</td>
</tr>
</tbody>
</table>

In [20], Feldman et al. had analyzed the maximal association rules between two item sets $X$ and $Y$, $X \Rightarrow Y$. The studies by Feldman [2] were based on the notion that whenever $X$ occurs alone then $Y$ also appears, with certain confidence. For a transactional $t$ in the database $D = \{ t_1, t_2, \cdots, t_m \}$, a category $T_i$ and an item set $X \subseteq T_i$, we say that $X$ is alone in $t$ if $t \cap T_i = X$. Formally, a maximal association rule implies that for $X \Rightarrow Y$, where $X$ and $Y$ subsets distinct categories, $T(X)$ and $T(Y)$, respectively. The support of the maximal association rule $X \Rightarrow Y$, denoted by $\sup_D^\max \left( X \Rightarrow Y \right)$, is defined as

$$\sup_D^\max \left( X \Rightarrow Y \right) = \left| \left\{ t : t \text{ maximal supports } X \text{ and } \left| t \text{ supports } Y \right| \right\} \right|.$$

The confidence of the maximal association rule $X \Rightarrow Y$, denoted by $\conf_D^\max \left( X \Rightarrow Y \right)$, is defined as $\conf_D^\max \left( X \Rightarrow Y \right) = \frac{s_D^\max(X \Rightarrow Y)}{|D(X, T(Y))|}$, where $D(X, T(Y))$ is the subset of the database $D$ whereby the database contains all the transactions that maximally support $X$ and contain at least one element of $T(Y)$ (the category of $Y$).
**Proposition 7.** Every maximal association rule is a regular association.

Notice that, the notion of alone in [20] can be easily obtained from splitting information system.

**Definition 8.** Let \( S = (U, A, V_{[0,1]}, f) \) be an information system, representing a transactional database \( D = \{t_1, t_2, \ldots, t_m\} \). The maximal splitting of \( S = (U, A, V_{[0,1]}, f) \) is defined as sub-information systems \( S' = (U', A, V_{[0,1]}, f) \), where \( U' = X' \) is appears alone in set of transactions \( t \subseteq D \).

Based on the definition of maximal splitting, from Table 1, we have the following 2 sub-maximal table as shown in Table 3 and Table 4 below.

<table>
<thead>
<tr>
<th>( U'/A )</th>
<th>US</th>
<th>Can</th>
<th>Fra</th>
<th>Corn</th>
<th>Fish</th>
</tr>
</thead>
<tbody>
<tr>
<td>( u_1 )</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>( u_2 )</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>( \ldots )</td>
<td>( \ldots )</td>
<td>( \ldots )</td>
<td>( \ldots )</td>
<td>( \ldots )</td>
<td>( \ldots )</td>
</tr>
<tr>
<td>( u_9 )</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>( u_{10} )</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>( \ldots )</td>
<td>( \ldots )</td>
<td>( \ldots )</td>
<td>( \ldots )</td>
<td>( \ldots )</td>
<td>( \ldots )</td>
</tr>
</tbody>
</table>

**Table 3.** The first table.

<table>
<thead>
<tr>
<th>( U'/A )</th>
<th>US</th>
<th>Can</th>
<th>Fra</th>
<th>Corn</th>
<th>Fish</th>
</tr>
</thead>
<tbody>
<tr>
<td>( u_{11} )</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>( u_{12} )</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>( \ldots )</td>
<td>( \ldots )</td>
<td>( \ldots )</td>
<td>( \ldots )</td>
<td>( \ldots )</td>
<td>( \ldots )</td>
</tr>
<tr>
<td>( u_{29} )</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>( u_{30} )</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>( \ldots )</td>
<td>( \ldots )</td>
<td>( \ldots )</td>
<td>( \ldots )</td>
<td>( \ldots )</td>
<td>( \ldots )</td>
</tr>
</tbody>
</table>

**Table 4.** The second table.

From Tables 1, we can also obtain maximal splitting using parameterization of attributes. In this case, we have the following parameters

\[
E_1 = \{\text{USA, Canada}\}, \\
E_2 = \{\text{USA, Canada, France}\}, \\
E_3 = \{\text{Corn}\}, \\
E_4 = \{\text{Fish}\}
\]

**Figure 1.** Parameterization of attributes from Tables 3 and 4.
Then, based on parameterization of attributes in Figure 1, we can split the data in Table 1 into 2 sub information systems as shown in Tables 3 and 4. The structures of two maximal splitting are equivalent. Furthermore, we have the associated soft set as in Figures 2 and 3, respectively.

\[(F, A_1) = \{ \text{US} = \{u_1, \ldots, u_{10}\}, \text{Canada} = \{u_{11}, \ldots, u_{10}\}, \text{Corn} = \{u_{1}, \ldots, u_{10}\} \}\]

Figure 2. Soft set representing Table 3.

\[(F, A_2) = \{ \text{US} = \{u_{11}, \ldots, u_{30}\}, \text{Canada} = \{u_{11}, \ldots, u_{30}\}, \text{France} = \{u_{11}, \ldots, u_{30}\}, \text{Fish} = \{u_{11}, \ldots, u_{30}\} \}\]

Figure 3. Soft set representing Table 4.

For comparison, with regular association rules, we have the following rules captured from Table 1.

\[\{\text{USA, Canada}\} \Rightarrow \{\text{Corn}\}, \text{ with sup = conf} = 0.333\]
\[\{\text{USA, Canada}\} \Rightarrow \{\text{Fish}\}, \text{ with sup = conf} = 0.666\]
\[\{\text{USA, Canada, France}\} \Rightarrow \{\text{Fish}\}, \text{ with sup = conf} = 1\]

Figure 4. Association rules captured from Table 1.

However, for maximal association rules, we have the following rules captured from Tables 3 and 4.

\[\{\text{USA, Canada}\} \Rightarrow \{\text{Corn}\}, \text{ with sup = conf} = 1\]
\[\{\text{USA, Canada, France}\} \Rightarrow \{\text{Fish}\}, \text{ with sup = conf} = 1\]

Figure 5. Maximal association rules captured from Tables 3 and 4.

From Figure 2, we have the value \(F(E_1) = \{1, \ldots, 10\}\) and \(F(E_2) = \{1, \ldots, 10\}\). Since \(|F(E_1)| \cdot |F(E_2)| = 100\), then \(\{\text{USA, Canada}\} \Rightarrow \{\text{Corn}\}\). From Figure 3, we have the value \(F(E_2) = \{11, \ldots, 30\}\) and \(F(E_4) = \{11, \ldots, 30\}\). Since \(|F(E_2)| \cdot |F(E_4)| = 200\), then \(\{\text{USA, Canada, France}\} \Rightarrow \{\text{Fish}\}\). Those total dependencies are similar to the maximal association rules in Figure 5.
5.2 Decision making

Different with the previous application, that we can make a decision based on a decision table, where the condition and decision attributes are already given. In this sub-section we present another application of the attributes dependencies for decision making under multi-valued information system without pre-defined decision. In [1], at the beginning, rough set theory was used mainly to make a decision in term of object classification, i.e., to assign them to classes known a priori [2]. Based on the result of [12], we present the application of soft set theory for decision making. Roughly speaking, decision making is a form of object classification where we do not know which one is the decision class.

The process for selecting a decision using soft set theory based attributes dependency can be depicted in the following figure.

![Figure 6. The concept of decision making using attributes dependencies.](image)

The maximal degree of dependencies among attributes in an information system is used to select the best decision. The justification that the attributes dependency can be used to find crisper decision is given in the following proposition.

**Proposition 9.** Let \( S = (U, A, V, f) \) be an information system and let \( D \) and \( C \) be any subsets of \( A \). For a condition of \( D \) depends totally on \( C \), then \( \alpha_D(X) \leq \alpha_C(X) \), for every \( X \subseteq U \).

**Proof.** Let subsets of \( A \) in an information system \( S = (U, A, V, f) \) be \( C \) and \( D \). From the hypothesis, we have \( IND(C) \subseteq IND(D) \). Furthermore, the partitioning \( U / C \) is finer than that of \( U / D \), thus, it is clear that any equivalence class induced by \( IND(D) \) is a union of some equivalence classes induced by \( IND(C) \). Therefore, for every \( x \in X \subseteq U \), we have \( [x]_C \subseteq [x]_D \). And hence, for every \( X \subseteq U \), we have \( D(X) \subseteq C(X) \subseteq X \subseteq \overline{C(X)} \subseteq \overline{D(X)} \).

Consequently

\[
\alpha_D(X) = \frac{|D(X)|}{|D(X)|} \leq \frac{|C(X)|}{|C(X)|} = \alpha_C(X) \square
\]
We elaborate our approach through a categorical-valued information system derived from the animal data set from [21]. In Table 5, there are nine animals (|U| = 9) with nine (|A| = 9) categorical attributes (|V_a| ≥ 2): Hair, Teeth, Eye, Feather, Feet, Eat, Milk, Fly, and Swim. The attributes Hair, Eye, Feather, Milk, Fly, and Swim have two values. Attributes Teeth has three values, and other attributes have four values.

Table 5. The animal world data set from [21].

<table>
<thead>
<tr>
<th>Animal</th>
<th>Hair</th>
<th>Teeth</th>
<th>Eye</th>
<th>Feather</th>
<th>Feet</th>
<th>Eat</th>
<th>Milk</th>
<th>Fly</th>
<th>Swim</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Tiger</td>
<td>Y</td>
<td>Pointed</td>
<td>Forward</td>
<td>N</td>
<td>Claw</td>
<td>Meat</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
</tr>
<tr>
<td>2. Cheetah</td>
<td>Y</td>
<td>Pointed</td>
<td>Forward</td>
<td>N</td>
<td>Claw</td>
<td>Meat</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
</tr>
<tr>
<td>3. Giraffe</td>
<td>Y</td>
<td>Blunt</td>
<td>Side</td>
<td>N</td>
<td>Hoof</td>
<td>Grass</td>
<td>Y</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>4. Zebra</td>
<td>Y</td>
<td>Blunt</td>
<td>Side</td>
<td>N</td>
<td>Hoof</td>
<td>Grass</td>
<td>Y</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>5. Ostrich</td>
<td>N</td>
<td>N</td>
<td>Side</td>
<td>Y</td>
<td>Claw</td>
<td>Grain</td>
<td>N</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>6. Penguin</td>
<td>N</td>
<td>N</td>
<td>Side</td>
<td>Y</td>
<td>Web</td>
<td>Fish</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>7. Albatross</td>
<td>N</td>
<td>N</td>
<td>Side</td>
<td>Y</td>
<td>Claw</td>
<td>Grain</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>8. Eagle</td>
<td>N</td>
<td>N</td>
<td>Forward</td>
<td>Y</td>
<td>Claw</td>
<td>Meat</td>
<td>N</td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>9. Viper</td>
<td>N</td>
<td>Pointed</td>
<td>Forward</td>
<td>N</td>
<td>N</td>
<td>Meat</td>
<td>N</td>
<td>N</td>
<td>N</td>
</tr>
</tbody>
</table>

The multi-soft sets representing Table 5 is given below.

\[
(F, E) = \left\{ \\begin{array}{l}
\{Y = 1,2,3,4\}, \{N = 5,6,7,8,9\}, \\
\{\text{Pointed} = 1,2,9\}, \{\text{Blunt} = 3,4\}, \{N = 5,6,7,8\}, \\
\{\text{Forward} = 1,2,8,9\}, \{N = 3,4,5,6,7\}, \\
\{\text{T1,2,3,4,9}\}, \{\text{Claw} = 1,2,5,7,8\}, \{\text{Hoff} = 3,4\}, \{\text{Web} = 6\}, \{N = 9\}, \\
\{\text{Meat} = 1,2,8,9\}, \{\text{Grass} = 3,4\}, \{\text{Grain} = 5,7\}, \{\text{Fish} = 6\}, \\
\{Y = 1,2,3,4\}, \{N = 5,6,7,8,9\}, \\
\{N = 1,2,3,4,5,6,9\}, \{N = 8,9\}, \\
\{Y = 1,2,6,7\}, \{N = 3,4,5,8,9\} \end{array} \right\}
\]

Figure 7. The multi-soft sets from Table 5.

Based on Figure 7, we use the formula in Definition 6 and obtain the attribute dependencies as shown in Figure 8. Based on Table 6, the first maximum degree of dependency of attributes, i.e. 1 appears in attributes Hair/Milk, Eye and Feather. Based on this technique, the next degree of attributes dependencies will be considered, until it is no longer tied. In this case, the second degree corresponding to attribute Hair/Milk, i.e. 0.666 is higher than that of attribute Eye and Feather, i.e. 0.555. Therefore, attribute Hair/Milk is selected as the best decision. After a decision attribute is selected, we can check the consistency of the decision system. Each object \( x \in U \) of a decision system determines a decision rule \( \text{AND}_{a \in A} \delta(x) \Rightarrow \delta(x) \). Decision rules are frequently defined as implications of “if...then...”. And for any set of decision rules, it is known as decision algorithm. The ratio of the number of consistent rules, i.e. sure rules (rules with no conflict) to all rules in a decision table is called consistency factor \( \gamma(C, D) \) of a decision table, denoted by \( \gamma(C, D) \). For Table 5, with attribute Hair/Milk as the decision attribute, we have \( \gamma(C, D) = 1 \). Thus the table is called a consistent decision system.
Figure 8. The attributes dependencies.

Table 6. The selecting attributes.

<table>
<thead>
<tr>
<th>Attributes</th>
<th>1st degree</th>
<th>2nd degree</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hair/Milk</td>
<td>1</td>
<td>0.666</td>
</tr>
<tr>
<td>Eye</td>
<td>1</td>
<td>0.555</td>
</tr>
<tr>
<td>Feather</td>
<td>1</td>
<td>0.555</td>
</tr>
</tbody>
</table>

5.3 Categorical data clustering

The last application of soft attributes dependencies is used for categorical data clustering. In this case, the main problem of clustering is how to select the clustering attribute among the candidates. We may use the result of selecting decision attribute in the above sub-section as the clustering attribute. From Table 5, the selected decision attribute will be the guidance in clustering (partition) the animals, i.e., Hair/Milk. As can be noted, the partition derived for the set of animals induced by attribute Hair/Milk is \{1,2,3,4\}, \{5,6,7,8,9\}. By which, we can split the animals based on the hierarchical tree as in Figure 9.

The recursive approach will be adapted in obtaining further clusters. At ensuing iterations, further splitting is done based on the leaf node having more objects. When it reaches a pre-defined number of clusters, the algorithm terminates. This is subjective and is pre-decided based on either user requirement or domain knowledge. For the future activities, we use the proposed technique for categorical data clustering through larger data set likes the benchmark datasets from some standard UCI database.
6. Conclusions

In this paper we have presented the applicability of soft set theory based on the attributes dependencies. By relying on fact that every rough set can be considered as a soft set, we successfully have discovered the attributes dependency in information systems based on the notion of mapping inclusion in the soft sets. The results achieved are comparable to the rough dependencies.

Acknowledgements. The authors would like to extend appreciation of the support given by Universiti Sultan Zainal Abidin (UniSZA), Terengganu Malaysia.

References


Discovering attributes dependency


http://dx.doi.org/10.1016/j.eswa.2011.04.011


http://dx.doi.org/10.1007/978-3-540-73451-2_13


http://dx.doi.org/10.1109/hicss.1995.375608


http://dx.doi.org/10.1109/ams.2009.148


http://dx.doi.org/10.1007/11004011_8


**Received: April 15, 2015; Published: November 2, 2015**