Rough Set Approach for Categorical Data Clustering[1]

Tutut Herawan*[1], Rozaida Ghazali[2], Iwan Tri Riyadi Yanto[3], and Mustafa Mat Deris[2]

[1]*Department of Mathematics Education*
Universitas Ahmad Dahlan, Yogyakarta, Indonesia
[2]*Faculty of Information Technology and Multimedia*
Universiti Tun Hussein Onn Malaysia, Johor, Malaysia
[3]*Department of Mathematics*
Universitas Ahmad Dahlan, Yogyakarta, Indonesia

tutut81@uad.ac.id (corresponding author),*
rozaida@uthm.edu.my, iwan015@gmail.com, mmustafa@uthm.edu.my

Abstract

Clustering categorical data is an integral part of data mining and has attracted much attention recently. In this paper, we focus our discussion on the rough set theory for categorical data clustering. We propose MADE (Maximal Attributes DEpendency), an alternative technique for categorical data clustering using rough set theory taking into account maximum attributes dependencies degree in categorical-valued information systems. Experimental results on two benchmark UCI datasets show that MADE technique is better with the baseline categorical data clustering technique with respect to computational complexity and clusters purity.

Keywords*: Clustering; Categorical data; Information system; Rough set theory; Attributes dependencies.*

1. Introduction

Clustering a set of objects into homogeneous classes is a fundamental operation in data mining. The operation is required in a number of data analysis tasks, such as unsupervised classification and data summation, as well as in the segmentation of large homogeneous datasets into smaller homogeneous subsets that can be easily managed, modeled separately and analyzed. Recently, many attentions have been paid on the categorical data clustering [1,2], where data objects are made up of non-numerical attributes. For categorical data clustering, several new trends have emerged for the techniques in handling uncertainty in the clustering process. One of the popular approaches for handling uncertainty is based on rough set theory [3]. The main idea of the rough clustering is the clustering dataset is mapped as the decision table. This can be done by introducing a decision attribute and consequently, a divide-and-conquer method can be used to partition/cluster the objects. The first attempt on rough set-based technique is to select clustering attribute proposed by Mazlack *et al.* [4]. They proposed two techniques, i.e., Bi-Clustering and TR techniques which are based on the bi-valued attribute and maximum total roughness in each attribute, respectively. One of the most successful pioneering rough clustering techniques is

[1] An early version of this paper appeared in the Proceeding of International Conference, DTA 2009, held as Part of the Future Generation Information Technology Conference, FGIT 2009, Jeju Island, Korea, December 10-12, 2009, CCIS 64 Springer-Verlag, pp. 179–186, 2009.

Minimum-Minimum Roughness (MMR) proposed by Parmar [5]. The technique is based on lower, upper and quality of approximations of a set [6]**.** However, since application of rough set theory in categorical data clustering is relatively new, the focus of MMR is still on the evaluation its performance. To this, the computational complexity and clusters purity are still outstanding issues since all attributes are considered for selection and objects in different class appear in a cluster, respectively.

In this paper, we propose MADE (Maximal Attributes DEpendency), an alternative technique for categorical data clustering. The technique differs on the baseline method, where the rough attributes dependencies in categorical-valued information systems is used to select clustering attribute based on the maximum degree. Further, we use a divide-and-conquer method to partition/cluster the objects. We have succeed in showing that the proposed technique is able to achieve lower computational complexity with higher purity as compared to MMR.

The rest of this paper is organized as follows. Section 2 describes rough set theory. Section 3 describes the analysis and comparison of Mazlack's TR and MMR techniques. Section 4 describes the Maximum Attributes Dependency (MADE) technique. Comparison tests of MADE with MMR techniques based on Soybean and Zoo datasets are described in section 5. Finally, the conclusion of this work is described in section 6.

2. Rough Set Theory

The syntax of information systems is very similar to relations in relational data bases. Entities in relational databases are also represented by tuples of attribute values. An *information system* is a 4-tuple (quadruple) $S=(U,A,V,f)$, where $U=\{u_1,u_2,u_3,\cdots,u_{|U|}\}$ is a non-empty finite set of objects, $A=\{a_1,a_2,a_3,\cdots,a_{|A|}\}$ 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. An information system is also called a knowledge representation systems or an attribute-valued system and can be intuitively expressed in terms of an information table (see Table 1).

Table 1. An information system

U	a_1	a_2	…	a_k	…	$a_{\|A\|}$
u_1	$f(u_1,a_1)$	$f(u_1,a_2)$	…	$f(u_1,a_k)$	…	$f(u_1,a_{\|A\|})$
u_2	$f(u_2,a_1)$	$f(u_2,a_2)$	…	$f(u_2,a_k)$	…	$f(u_2,a_{\|A\|})$
⋮	⋮	⋮	⋱	⋮	⋱	⋮
$u_{\|U\|}$	$f(u_{\|U\|},a_1)$	$f(u_{\|U\|},a_2)$	…	$f(u_{\|U\|},a_k)$	…	$f(u_{\|U\|},a_{\|A\|})$

The time complexity for computing an information system $S=(U,A,V,f)$ is $|U|\times|A|$ since there are $|U|\times|A|$ values of $f(u_i,a_j)$ to be computed, where $i=1,2,3,\cdots,|U|$ and $j=1,2,3,\cdots,|A|$. Note that *t* induces a set of maps $t=f(u,a):U\times A\to V$. Each map is a tuple $t_i=(f(u_i,a_1),f(u_i,a_2),f(u_i,a_3),\cdots,f(u_i,a_{|A|}))$, where where $i=1,2,3,\cdots,|U|$. Note that the tuple *t* is not necessarily associated with entity uniquely (see Table 7). In an information table, two distinct entities could have the same tuple representation

(duplicated/redundant tuple), which is *not permissible* in relational databases. Thus, the concept of information systems is a generalization of the concept of relational databases.

Definition 1. *Two elements* $x, y \in U$ *are 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* $a \in B$.

Obviously, every subset of *A* induces unique indiscernibility relation. Notice that, an indiscernibility relation induced by the set of attribute *B*, denoted by $IND(B)$, is an equivalence relation. The partition of *U* induced by $IND(B)$ is denoted by U/B and the equivalence class in the partition U/B containing $x \in U$, is denoted by $[x]_B$. The notions of lower and upper approximations of a set are defined as follows.

Definition 2. (See [6].) *The B-lower approximation of X, denoted by* $\underline{B}(X)$ *and B-upper approximations of X, denoted by* $\overline{B}(X)$, *are defined by*

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

It is easily seen that the upper approximation of a subset $X \subseteq U$ is expressed using set complement and lower approximation by

$$\overline{B}(X) = U - \underline{B}(\neg X),$$

where $\neg X$ denote the complement of *X* relative to *U*.

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

$$\alpha_B(X) = |\underline{B}(X)| / |\overline{B}(X)|, \qquad (1)$$

where $|X|$ denotes the cardinality of *X*. For empty set ϕ, we define $\alpha_B(\phi) = 1$. Obviously, $0 \le \alpha_B(X) \le 1$. If *X* is a union of some equivalence classes, then $\alpha_B(X) = 1$. Thus, the set *X* is *crisp* with respect to *B*, and otherwise, if $\alpha_B(X) < 1$, *X* is *rough* with respect to *B*.

The accuracy of roughness in equation (1) can also be interpreted using the well-known Marczeweski-Steinhaus (MZ) metric [7]. By applying the Marczeweski-Steinhaus metric to the lower and upper approximations of a subset $X \subseteq U$ in information system *S*, we have

$$D(\underline{B}(X), \overline{B}(X)) = 1 - \frac{|\underline{B}(X) \cap \overline{B}(X)|}{|\underline{B}(X) \cup \overline{B}(X)|} = 1 - \frac{|\underline{B}(X)|}{|\overline{B}(X)|} = 1 - \alpha_B(X). \qquad (2)$$

The notion of the dependency of attributes in information systems is given in the following definition.

Definition 3. *Let* $S = (U, A, V, f)$ *be an information system and let D and C be any subsets of A. Attribute D is called depends totally on attribute C, denoted* $C \Rightarrow D$, *if all values of attributes D are uniquely determined by values of attributes C.*

In other words, attribute *D* depends totally on attribute *C*, if there exist a functional dependency between values *D* and *C*. The notion of generalized attributes dependency is given in the following definition.

Definition 4. *Let* $S=(U, A, V, f)$ *be an information system and let D and C be any subsets of A. Degree of dependency of attribute D on attributes C, denoted* $C \Rightarrow_k D$*, is defined by*

$$k=\frac{\sum_{X \in U / D}|\underline{C}(X)|}{|U|}. \tag{3}$$

Obviously, $0 \le k \le 1$. Attribute *D* is said to be (totally dependent) depends totally (in a degree of *k*) on the attribute *C* if $k=1$. Otherwise, *D* is depends partially on *C*. Thus, attribute *D* depends totally (partially) on attribute *C*, if all (some) elements of the universe *U* can be uniquely classified to equivalence classes of the partition U / D, employing *C*.

In the following section, we analyze and compare the Total Roughness (TR) and Min-Min Roughness (MMR) techniques for selecting a clustering attribute.

3. TR and MMR Techniques

3.1. The TR Technique

The definition of information system is based on the notion of information system as stated in section 2. From the definition, suppose that attribute $a_i \in A$ has *k*-different values, say β_k, $k=1,2,\cdots,n$. Let $X(a_i=\beta_k)$, $k=1,2,\cdots,n$ be a subset of the objects having *k*-different values of attribute a_i. The roughness of TR technique of the set $X(a_i=\beta_k)$, $k=1,2,\cdots,n$, with respect to a_j, where $i \ne j$, denoted by $\mathrm{R}_{a_j}(X|a_i=\beta_k)$, is defined by

$$\mathrm{R}_{a_j}(X|a_i=\beta_k)=\frac{\left|\underline{X_{a_j}(a_i=\beta_k)}\right|}{\left|\overline{X_{a_j}(a_i=\beta_k)}\right|},\ k=1,2,\cdots,n. \tag{4}$$

From TR technique, the mean roughness of attribute $a_i \in A$ with respect to attribute $a_j \in A$, where $i \ne j$, denoted $\mathrm{Rough}_{a_j}(a_i)$, is evaluated as follow

$$\mathrm{Rough}_{a_j}(a_i)=\frac{\sum_{k=1}^{|V(a_i)|}\mathrm{R}_{a_j}(X|a_i=\beta_k)}{|V(a_i)|}, \tag{5}$$

where $V(a_i)$ is the set of values of attribute $a_i \in A$.

The total roughness of attribute $a_i \in A$ with respect to attribute $a_j \in A$, where $i \ne j$, denoted $\mathrm{TR}(a_i)$, is obtained by the following formula

$$\mathrm{TR}(a_i) = \frac{\sum_{j=1}^{|A|} \mathrm{Rough}_{a_j}(a_i)}{|A| - 1}. \quad (6)$$

As stated in Mazlack *et al.* [4], the highest value of TR, is the best selection of partitioning attribute.

3.2. The MMR Technique

The definition of information system is based on the notion of information system as stated in section 2. From the definition, suppose that attribute $a_i \in A$ has *k*-different values, say β_k, $k = 1,2,\cdots,n$. Let $X(a_i = \beta_k)$, $k = 1,2,\cdots,n$ be a subset of the objects having *k*-different values of attribute a_i. The roughness of MMR technique of the set $X(a_i = \beta_k)$, $k = 1,2,\cdots,n$, with respect to a_j, where $i \neq j$, denoted by $\mathrm{R}_{a_j}(X|a_i = \beta_k)$, is defined by

$$\mathrm{MMR}_{a_j}(X|a_i = \beta_k) = 1 - \frac{\left|\underline{X}_{a_j}(a_i = \beta_k)\right|}{\left|\overline{X}_{a_j}(a_i = \beta_k)\right|},\ k = 1,2,\cdots,n. \quad (7)$$

It is clear that MMR technique uses MZ metric to measure the roughness of the set $X(a_i = \beta_k)$, $k = 1,2,\cdots,n$, with respect to a_j, where $i \neq j$.

The mean roughness of MMR technique is defined by

$$\mathrm{MMRough}_{a_j}(a_i) = \frac{\sum_{k=1}^{|V(a_i)|} \mathrm{MMR}_{a_j}(X|a_i = \beta_k)}{|V(a_i)|}. \quad (8)$$

According to Parmar *et al.* [5], the least mean roughness is the best selection of partitioning attribute.

3.3. Comparison of TR and MMR techniques

Proposition 5. *The value of roughness of MMR technique is the opposite of that TR technique.*

Proof. Since MMR technique uses MZ metric to measure the roughness of the set $X(a_i = \beta_k)$, $k = 1,2,\cdots,n$, with respect to a_j, where $i \neq j$, i.e.,

$$\mathrm{MMR}_{a_j}(X|a_i = \beta_k) = 1 - \frac{\left|\underline{X}_{a_j}(a_i = \beta_k)\right|}{\left|\overline{X}_{a_j}(a_i = \beta_k)\right|},$$

then from (7), we have

$$\mathrm{MMR}_{a_j}(X|a_i = \beta_k) = 1 - \mathrm{R}_{a_j}(X|a_i = \beta_k). \quad (9)$$

Thus, the value of mean roughness of MMR technique is also the opposite of that TR technique (5), i.e.,

$$\text{MMRough}_{a_j}(a_i) = \frac{\sum_{k=1}^{|V(a_i)|} \text{MMR}_{a_j}(X|a_i = \beta_k)}{|V(a_i)|}$$

$$= \frac{\sum_{k=1}^{|V(a_i)|} \left(1 - \text{R}_{a_j}(X|a_i = \beta_k)\right)}{|V(a_i)|}$$

$$= \frac{\sum_{k=1}^{|V(a_i)|} 1 - \sum_{k=1}^{|V(a_i)|} \text{R}_{a_j}(X|a_i = \beta_k)}{|V(a_i)|}$$

$$= \frac{|V(a_i)|}{|V(a_i)|} - \frac{\sum_{k=1}^{|V(a_i)|} \text{R}_{a_j}(X|a_i = \beta_k)}{|V(a_i)|}$$

$$= 1 - \text{Rough}_{a_j}(a_i), \text{ for } i \neq j. \quad (10)$$

The MMR technique is based on the minimum value of mean roughness in (10), without calculating total roughness (6).

This analysis and comparison has shown that TR and MMR techniques are providing the similar result when used in determining the clustering attribute. To illustrate that MMR and Mazlack's techniques provide the same results, we consider to the following example.

Example 6. We consider the dataset in illustrative example of Table 2 in [5].

Table 2. An information system in [5]

U	a_1	a_2	a_3	a_4	a_5	a_6
1	Big	Blue	Hard	Indefinite	Plastic	Negative
2	Medium	Red	Moderate	Smooth	Wood	Neutral
3	Small	Yellow	Soft	Fuzzy	Plush	Positive
4	Medium	Blue	Moderate	Fuzzy	Plastic	Negative
5	Small	Yellow	Soft	Indefinite	Plastic	Neutral
6	Big	Green	Hard	Smooth	Wood	Positive
7	Small	Yellow	Hard	Indefinite	Metal	Positive
8	Small	Yellow	Soft	Indefinite	Plastic	Positive
9	Big	Green	Hard	Smooth	Wood	Neutral
10	Medium	Green	Moderate	Smooth	Plastic	Neutral

In Table 2, there are ten objects $(|U| = 10)$ with six categorical-valued attributes: a_1, a_2, a_3, a_4, a_5 and a_6. Each attribute has more than two values $(Va_i > 2)$, $i = 1,2,3,4,5,6$. Since in this case there is no bi-valued attributes, then we cannot employ Mazlack's BC technique. The calculation of TR and MMR techniques must be applied on all of the attribute values for obtaining the clustering attribute. The calculation of TR value is based on formulas in (4), (5) and (6).

The techniques of TR and MMR are implemented in MATLAB version 7.6.0.324 (R2008a). They are executed sequentially on a processor Intel Core 2 Duo CPUs. The total

main memory is 1G and the operating system is Windows XP Professional SP3. The results of TR and MMR are given in the following Table 3 and 4, respectively.

Table 3. The TR of all attributes of Table 2

Attribute	**TR mean roughness**				
a_1	Rough a_2	Rough a_3	Rough	Rough	Rough
	0.3889	0.4762	0	0.0476	0
a_2	Rough a_1	Rough a_3	Rough	Rough	Rough
	0.2500	0.1071	0	0.0357	0.2500
a_3	Rough a_1	Rough a_2	Rough	Rough	Rough
	0.4762	0.0556	0	0.0333	0
a_4	Rough a_1	Rough a_2	Rough	Rough	Rough
	0	0.3333	0	0.1587	0
a_5	Rough a_1	Rough a_2	Rough	Rough	Rough
	0	0.1574	0.1000	0.0667	0.0667
a_6	Rough a_1	Rough a_2	Rough	Rough	Rough
	0	0.3750	0	0	0.0333

Table 4. The MMR of all attributes of Table 2

Attribute	**MMR mean roughness**				
a_1	Rough a_2	Rough a_3	Rough a_4	Rough a_5	Rough a_6
	0.6111	0.5238	1	0.9048	1
a_2	Rough a_1	Rough a_3	Rough a_4	Rough a_5	Rough a_6
	0.7500	0.8929	1	0.9286	0.7500
a_3	Rough a_1	Rough a_2	Rough a_4	Rough a_5	Rough a_6
	0.5238	0.9444	1	0.9074	1
a_4	Rough a_1	Rough a_2	Rough a_3	Rough a_5	Rough a_6
	1	0.6667	1	0.7639	1
a_5	Rough a_1	Rough a_2	Rough a_3	Rough a_4	Rough a_6
	1	0.8820	1	1	0.9500
a_6	Rough a_1	Rough a_2	Rough a_3	Rough a_4	Rough a_5
	1	0.6250	1	1	0.9333

Based on Figure 1, attribute a_1, i.e., 0.1825 has higher TR as compared to a_i, $i = 2,3,4,5,6$. Thus, attribute a_1 is selected as the clustering attribute. Meanwhile, based on Figure 2, two attributes are of equally of MMR (a_1 and a_3 , i.e. 0.5238). But, the second value corresponding to attribute a_1, i.e. 0.6111 is lower than that of a_3 , i.e. 0.9074. Therefore, attribute a_1 is selected as the clustering attribute.

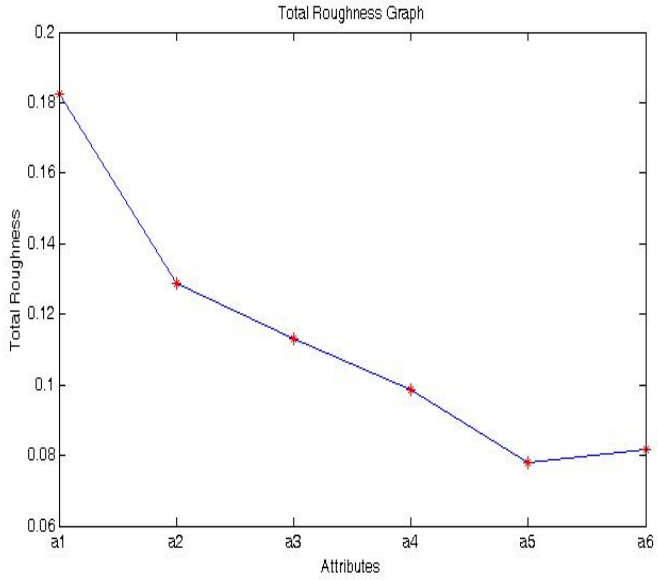

Figure 1. The TR value of all attributes of Table 2

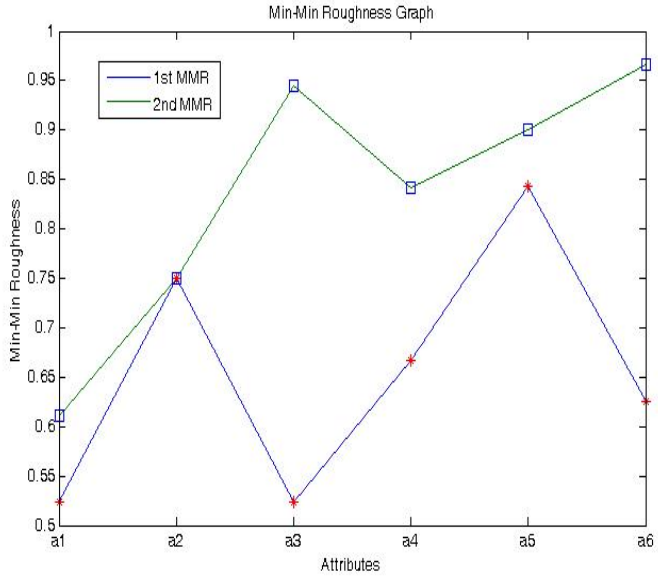

Figure 2. The MMR value of all attributes of Table 2

Table 5. The computation and response time of TR and MMR

	Computation	Response time (Sec)
TR	237	0.047
MMR	237	0.047

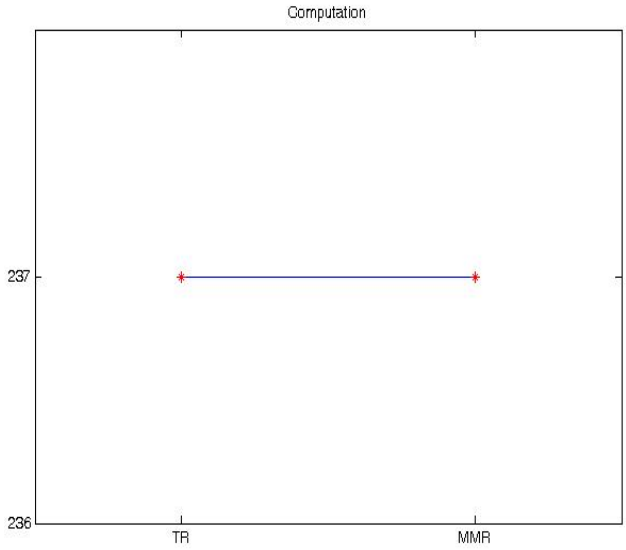

Figure 3. The computation of TR and MMR

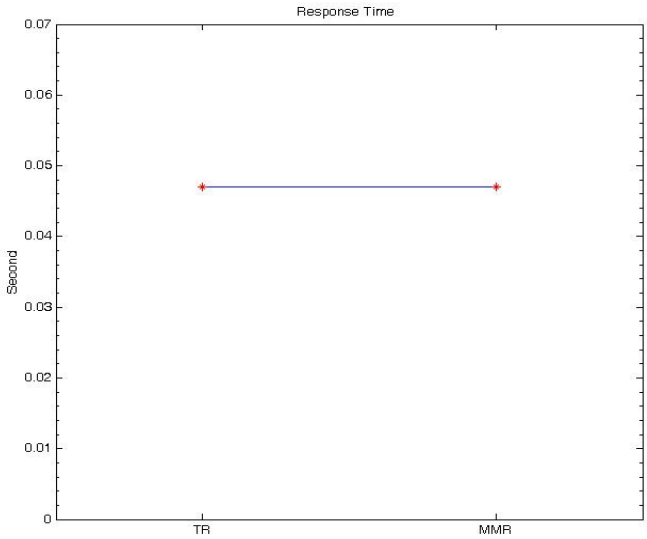

Figure 4. The response time of TR and MMR

Based on the result on selecting clustering attribute in Figures 1, 2, 3 and 4, it is easily seen that the decision, computation complexity and processing time of TR and MMR techniques are totally the same. Thus based on Proposition 5, the statement that MMR is an extension of an approach proposed Mazlack *et al.* in comparison example [5] is therefore considered as incorrect and unreasonable. On the other hand, to achieve lower computational complexity in selecting partitioning attribute using MMR, Parmar *et al.* suggested that the measurement of the roughness to be based on relationship between an attribute $a_i \in A$ and the set defined as $A - \{a_i\}$ instead of calculating the maximum with respect to all $\{a_j\}$ where $a_i \neq a_j$ [5]. As has been observed by us, this technique only can be applied to a very special dataset. To illustrate this problem, we consider to the following example.

Example 7. In Table 2, if we consider to measure the roughness of attribute $a_i \in A$ with respect to the set of attributes $A - \{a_i\}$, then we get the value of modified MMR as in Table 6.

Table 6. The modified MMR of all attributes of dataset in [5]

Attribute w.r.t.	Mean Roughness	MMR
a_1	Rough $A-\{a_1\}$ 0	0
a_2	Rough $A-\{a_2\}$ 0	0
a_3	Rough $A-\{a_3\}$ 0	0
a_4	Rough $A-\{a_4\}$ 0	0
a_5	Rough $A-\{a_5\}$ 0	0
a_6	Rough $A-\{a_6\}$ 0	0

Based on Table 6, we have not been able to select a clustering attribute. Thus, the suggested technique would lead a problem, i.e., after calculation of mean roughness of attribute $a_i \in A$ with respect to the set of attributes $A - \{a_i\}$, the value of MMR usually cannot preserve the original decision. Thus, this modified technique is therefore not relevant to all type of dataset.

To overcome the problem of computational complexity of MMR, in section 4, we introduce the Maximum Attributes Dependencies (MADE) technique to deal with the problem of categorical data clustering.

4. Maximum Attributes DEpendencies (MADE) Technique

4.1. MADE technique

The MADE technique for selecting partitioning attribute is based on the maximum degree of dependency of attributes. The justification that the higher of the degree of dependency of attributes implies the more accuracy for selecting partitioning attribute is stated in the Proposition 8.

Proposition 8. *Let $S = (U, A, V, f)$ be an information system and let D and C be any subsets of A. If D depends totally on C, then*

$$\alpha_D(X) \leq \alpha_C(X),$$

for every $X \subseteq U$.

Proof. Let *D* and *C* be any subsets of *A* in information system $S = (U, A, V, f)$. From the hypothesis, we have $IND(C) \subseteq IND(D)$. Furthermore, the partitioning U/C is finer than that U/D, thus, it is clear that any equivalence class induced by $IND(D)$ is a union of

some equivalence class 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

$$\underline{D}(X) \subseteq \underline{C}(X) \subset X \subset \overline{C}(X) \subseteq \overline{D}(X).$$

Consequently

$$\alpha_D(X) = \frac{|\underline{D}(X)|}{|\overline{D}(X)|} \leq \frac{|\underline{C}(X)|}{|\overline{C}(X)|} = \alpha_C(X). \square$$

4.2. Complexity

Suppose that in an information system $S = (U, A, V, f)$, there is $|A|$ attributes. For MADE, the computation of calculating of dependency degree of attribute a_i on attribute a_j, where $i \neq j$ is $|A| \times |A-1|$. Thus, the computational complexity for MADE technique is of the polynomial $O(|A| \times |A-1|)$.

The MADE's algorithm for selecting clustering attribute is given in Figure 5.

```
Algorithm: MADE
Input: Dataset without clustering attribute
Output: Clustering attribute
Begin
    Step 1. Compute the equivalence classes
            using the indiscernibility relation
            on each attribute.
    Step 2. Determine the dependency degree of
            attribute a_i with respect to all a_j,
            where i ≠ j.
    Step 3. Select the maximum of dependency
            degree of each attribute.
    Step 4. Select a clustering attribute based
            on the maximum degree of dependency
            of attributes.
End
```

Figure 5. The MADE algorithm

As the same procedure for selecting clustering attribute of MMR, in using MADE technique, it is recommended to look at the next lowest dependencies degree inside the attributes that are tied and so on until the tie is broken.

4.3. Example

The dataset is an animal dataset from Hu [8]. In Table 7, there are nine animals $(|U| = 9)$ with nine categorical-valued attributes $(|A| = 9)$; 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.

a. To obtain the dependencies degree of all attributes, the first step of the techniques is to obtain the equivalence classes induced by indiscernibility relation of singleton attributes, i.e., disjoint classes of objects which are contain indiscernible objects.
b. By collecting the equivalence classes, a partition of objects can be obtained. The partitions are shown in Figure 6.
c. The dependency degree of attributes can be obtained using formula in (3). For attribute Hair depends on attributes Teeth, Eye, Feather, Feet, Eat, Milk, Fly and Swim, we have the degrees as shown in Figure 7.

Table 7. Animal world dataset from [8]

Animal	Hair	Teeth	Eye	Feather	Feet	Eat	Milk	Fly	Swim
Tiger	Y	Pointed	Forward	N	Claw	Meat	Y	N	Y
Cheetah	Y	Pointed	Forward	N	Claw	Meat	Y	N	Y
Giraffe	Y	Blunt	Side	N	Hoof	Grass	Y	N	N
Zebra	Y	Blunt	Side	N	Hoof	Grass	Y	N	N
Ostrich	N	N	Side	Y	Claw	Grain	N	N	N
Penguin	N	N	Side	Y	Web	Fish	N	N	Y
Albatross	N	N	Side	Y	Claw	Grain	N	Y	Y
Eagle	N	N	Forward	Y	Claw	Meat	N	Y	N
Viper	N	Pointed	Forward	N	N	Meat	N	N	N

a. $X(\text{Hair} = \text{yes}) = \{1,2,3,4\}$, $X(\text{Hair} = \text{no}) = \{5,6,7,8,9\}$,
$U / \text{Hair} = \{\{1,2,3,4\}, \{5,6,7,8,9\}\}$.
b. $X(\text{Teeth} = \text{pointed}) = \{1,2,9\}$, $X(\text{Teeth} = \text{blunt}) = \{3,4\}$,
$X(\text{Teeth} = \text{no}) = \{5,6,7,8\}$,
$U / \text{Teeth} = \{\{1,2,9\}, \{3,4\}\{5,6,7,8\}\}$.
c. $X(\text{Eye} = \text{Forward}) = \{1,2,8,9\}$, $X(\text{Eye} = \text{Side}) = \{3,4,5,6,7\}$,
$U / \text{Eye} = \{\{1,2,8,9\}, \{3,4,5,6,7\}\}$.
d. $X(\text{Feather} = \text{no}) = \{1,2,3,4,9\}$, $X(\text{Feather} = \text{yes}) = \{5,6,7,8\}$,
$U / \text{Feather} = \{\{1,2,3,4,9\}, \{5,6,7,8\}\}$.
e. $X(\text{Feet} = \text{claw}) = \{1,2,5,7,8\}$, $X(\text{Feet} = \text{hoof}) = \{3,4\}$,
$X(\text{Feet} = \text{web}) = \{6\}$, $X(\text{Feet} = \text{no}) = \{9\}$.
$U / \text{Feet} = \{\{1,2,5,7,8,9\}, \{3,4\}, \{6\}, \{9\}\}$.
f. $X(\text{Eat} = \text{Meat}) = \{1,2,8,9\}$, $X(\text{Eat} = \text{grass}) = \{3,4\}$,
$X(\text{Eat} = \text{grain}) = \{5,7\}$, $X(\text{Eat} = \text{fish}) = \{6\}$.
$U / \text{Eat} = \{\{1,2,8,9\}, \{3,4\}, \{5,7\}, \{6\}\}$.
g. $X(\text{Milk} = \text{yes}) = \{1,2,3,4\}$, $X(\text{Milk} = \text{no}) = \{5,6,7,8,9\}$,
$U / \text{Milk} = \{\{1,2,3,4\}, \{5,6,7,8,9\}\}$.
h. $X(\text{Fly} = \text{no}) = \{1,2,3,4,5,6,9\}$, $X(\text{Fly} = \text{yes}) = \{7,8\}$,
$U / \text{Fly} = \{\{1,2,3,4,5,6\}, \{7,8\}\}$.
i. $X(\text{Swim} = \text{yes}) = \{1,2,6,7\}$, $X(\text{Swim} = \text{no}) = \{3,4,5,8,9\}$,
$U / \text{Swim} = \{\{1,2,6,7\}, \{3,4,5,8,9\}\}$.

Figure 6. The partitions using singleton attributes

$$\text{Teeth} \Rightarrow_k \text{Hair}, \text{ where } k = \frac{\sum_{U/\text{Hair}} |\underline{\text{Teeth}}(X)|}{|U|} = \frac{|\{3,4\}| + |\{5,6,7,8\}|}{9} = \frac{6}{9}.$$

$$\text{Eye} \Rightarrow_k \text{Hair}, \text{ where } k = \frac{\sum_{U/\text{Hair}} |\underline{\text{Eye}}(X)|}{|U|} = \frac{|\phi|}{9} = 0.$$

$$\text{Feather} \Rightarrow_k \text{Hair}, \text{ where } k = \frac{\sum_{U/\text{Hair}} |\underline{\text{Feather}}(X)|}{|U|} = \frac{|\{5,6,7,8\}|}{9} = \frac{4}{9}.$$

$$\text{Feet} \Rightarrow_k \text{Hair}, \text{ where } k = \frac{\sum_{U/\text{Hair}} |\text{Feet}(X)|}{|U|} = \frac{|\{3,4\}| + |\{6\}| + |\{9\}|}{9} = \frac{4}{9}.$$

$$\text{Eat} \Rightarrow_k \text{Hair}, \text{ where } k = \frac{\sum_{U/\text{Hair}} |\text{Eat}(X)|}{|U|} = \frac{|\{3,4\}| + |\{5,7\}| + |\{6\}|}{9} = \frac{5}{9}.$$

$$\text{Milk} \Rightarrow_k \text{Hair}, \text{ where } k = \frac{\sum_{U/\text{Hair}} |\text{Milk}(X)|}{|U|} = \frac{|\{1,2,3,4\}| + |\{5,6,7,8,9\}|}{9} = 1.$$

$$\text{Fly} \Rightarrow_k \text{Hair}, \text{ where } k = \frac{\sum_{U/\text{Hair}} |\underline{\text{Fly}}(X)|}{|U|} = \frac{|\{7,8\}|}{9} = \frac{2}{9}.$$

$$\text{Swim} \Rightarrow_k \text{Hair}, \text{ where } k = \frac{\sum_{U/\text{Hair}} |\underline{\text{Swim}}(X)|}{|U|} = \frac{|\phi|}{9} = 0.$$

Figure 7. The attributes dependencies

Similar calculations are performed for all the attributes. These calculations are summarized in Table 8.

Table 8. The dependencies degree of all attributes from Table 7

Attribute	**Degree of dependency**							
Hair	Teeth 0.666	Eye 0	Feather 0.444	Feet 0.444	Eat 0.555	Milk 1	Fly 0.222	Swim 0
Teeth	Hair 0	Eye 0	Feather 0.444	Feet 0.444	Eat 0.555	Milk 0	Fly 0.222	Swim 0
Eye	Hair 0	Teeth 0.555	Feather 0	Feet 0.444	Eat 1	Milk 0	Fly 0	Swim 0
Feather	Hair 0.444	Teeth 1	Eye 0	Feet 0.444	Eat 0.555	Milk 0.444	Fly 0.222	Swim 0
Feet	Hair 0	Teeth 0.222	Eye 0	Feather 0	Eat 0.555	Milk 0	Fly 0.2222	Swim 0
Eat	Hair 0	Teeth 0.555	Eye 0.444	Feather 0	Feet 0.333	Milk 0	Fly 0	Swim 0
Milk	Hair 1	Teeth 0.666	Eye 0	Feather 0.444	Feet 0.444	Eat 0.555	Fly 0.222	Swim 0
Fly	Hair 0.44	Teeth 0.555	Eye 0	Feather 0.555	Feet 0.44	Eat 0.33	Milk 0.444	Swim 0

	4				4	3		
Swim	Hair	Teeth	Eye	Feather	Feet	Eat	Milk	Fly
	0	0.222	0	0	0.444	0.333	0	0

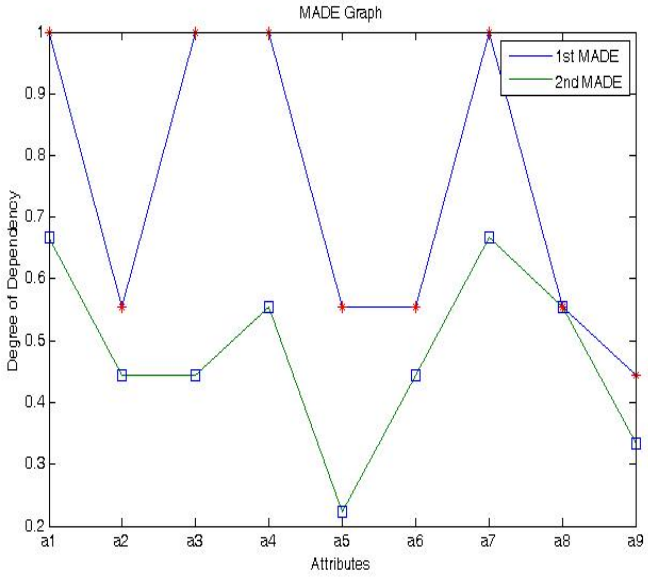

Figure 8. The maximal attributes dependencies

With the MADE technique, the first maximum degree of dependency of attributes, i.e. 1 occurs in attributes Hair (Milk), Eye and Feather (i.e., 1) as Figure 8 shows. The second maximum degree of dependency of attributes, i.e. 0.666 occurs in attributes Hair. Thus, based on Figure 8, attribute Hair is selected as clustering attribute.

4.4. Objects splitting

For objects splitting, we use a divide-conquer method. For example, in Table 7 we can cluster (partition) the animals based on the decision attribute selected, i.e., Hair/Milk. Notice that, the partition of the set of animals induced by attribute Hair/Milk is $\{\{1,2,3,4\},\{5,6,7,8,9\}\}$. To this, we can split the animals using the hierarchical tree as follows.

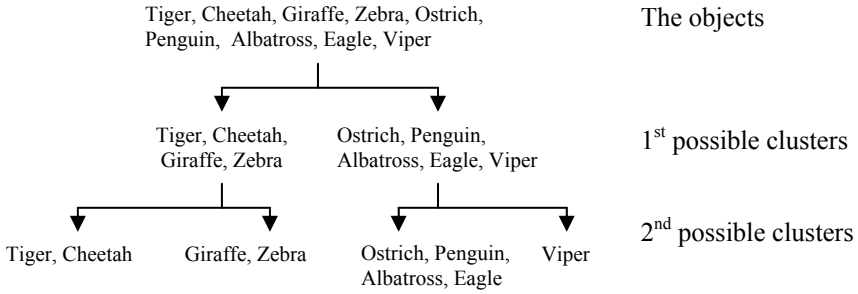

Figure 9. The objects splitting

The technique is applied recursively to obtain further clusters. At subsequent iterations, the leaf node having more objects is selected for further splitting. The algorithm terminates

when it reaches a pre-defined number of clusters. This is subjective and is pre-decided based either on user requirement or domain knowledge.

5. Comparison Tests

In order to test MADE and compare it with MMR, we use two datasets obtained from the benchmark UCI Machine Learning Repository. We use Soybean and Zoo datasets are with 47 and 101 objects. The purity of clusters was used as a measure to test the quality of the clusters [5]. The purity of a cluster and overall purity are defined as

$$\text{Purity}(i) = \frac{\text{the number of data occuring in both the } i\text{th cluster and its corresponding class}}{\text{the number of data in the data set}}$$

$$\text{Overall Purity} = \frac{\sum_{i=1}^{\#\text{ of cluster}} \text{Purity}(i)}{\#\text{ of cluster}}$$

According to this measure, a higher value of overall purity indicates a better clustering result, with perfect clustering yielding a value of 1 [5]. The algorithms of MMR and MADE for Soybean and Zoo datasets are implemented in MATLAB version 7.6.0.324 (R2008a). They are executed sequentially on a processor Intel Core 2 Duo CPUs. The total main memory is 1 Gigabyte and the operating system is Windows XP Professional SP3.

5.1. Soybean dataset

The Soybean dataset contains 47 objects on diseases in soybeans. Each object can be classified as one of the four diseases namely, Diaporthe Stem Canker (D1), Charcoal Rot (D2), Rhizoctonia Root Rot (D3), and Phytophthora Rot (D4) and are described by 35 categorical attributes [9]. The dataset is comprised 17 objects for Phytophthora Rot disease and 10 objects for each of the remaining diseases. Since there are four possible diseases, the objects will be split into four clusters. The results are summarized in Table 9. All of 47 objects belong to the majority class label of the cluster in which they are classified. Thus, the overall purity of the clusters is 100%.

Table 9. The purity of clusters

Cluster number	D1	D2	D3	D4	Purity
1	10	0	0	0	1
2	0	10	0	0	1
3	0	0	10	0	1
4	0	0	0	17	1
Overall Purity					1

5.2. Zoo dataset

The Zoo dataset is comprised of 101 objects, where each data point represents information of an animal in terms of 18 categorical attributes [10]. Each animal data point is classified into seven classes. Therefore, for MADE, the splitting data is set at seven clusters. Table 10 summarizes the results of running the MADE algorithm on the Zoo dataset.

Table 10. The purity of clusters

Cluster number	C1	C2	C3	C4	C5	C6	C7	Purity
1	41	0	0	0	0	0	0	1
2	0	20	0	0	0	0	0	1
3	0	0	5	0	0	0	0	1
4	0	0	0	13	0	0	0	1
5	0	0	0	0	4	0	0	1
6	0	0	0	0	0	8	0	1
7	0	0	0	0	0	0	10	1
Overall Purity								1

All of 101 objects belong to the majority class label of the cluster in which they are classified. Thus, the overall purity of the clusters is 100%.

5.3. Comparison

The comparison of overall purity, computation and response time of MADE and MMR on Soybean and Zoo datasets are given in Figures 10, 11 and 12, respectively. Based on Table 11, the MADE technique provides better solution compared to MMR technique both in Soybean and Zoo dataset.

Table 11. The overall improvement of MMR by MADE

	Improvement		
	Clusters Purity	Computation	Response Time
Soybean	17%	64%	63%
Zoo	9%	77%	67%

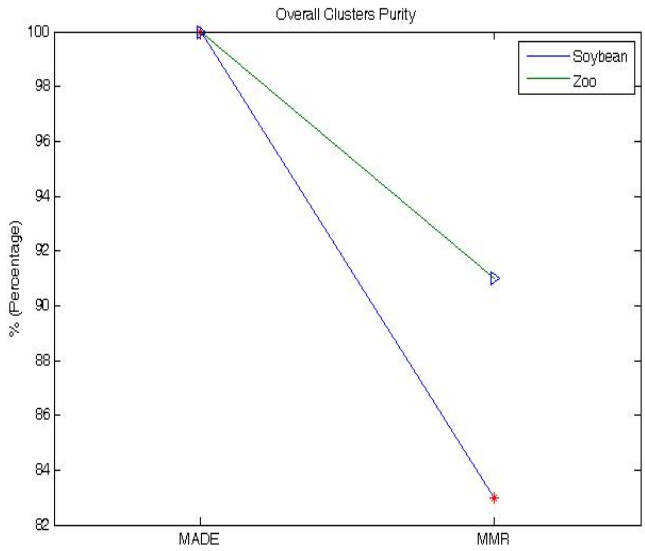

Figure 10. The comparison of overall purity

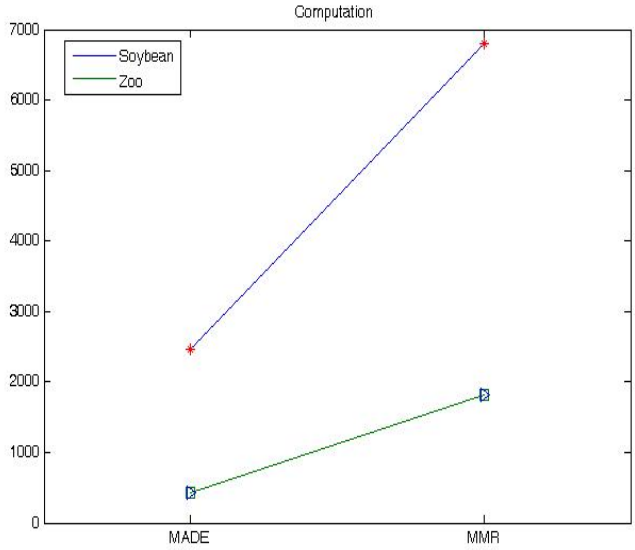

Figure 11. The comparison of computation

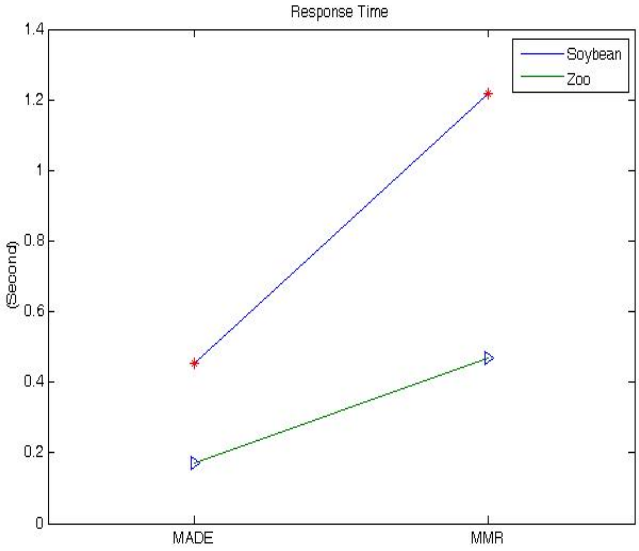

Figure 12. The comparison of response time

6. Conclusion

Categorical data clustering technique has emerged as a new trend in technique of handling uncertainty in the clustering process. In this paper, we have proposed MADE, an alternative technique for categorical data clustering using rough set theory based on attributes dependencies. We have proven that MADE technique is a generalization of MMR technique which is able to achieve lower computational complexity and higher clusters purity. With this approach, we believe that some applications through MADE will be applicable, such as for decision making, clustering very large datasets and etc.

Acknowledgement

This work was supported by the grant of Universiti Tun Hussein Onn Malaysia.

References

[1] Huang, Z. “Extensions to the k-means algorithm for clustering large data sets with categorical values”. Data Mining and Knowledge Discovery 2 (3) (1998) 283–304.
[2] Kim, D., Lee, K., Lee, D. “Fuzzy clustering of categorical data using fuzzy centroids”. Pattern Recognition Letters 25 (11) (2004) 1263–1271.
[3] Pawlak, Z. “Rough sets”. International Journal of Computer and Information Science. 11, 1982, 341–356.
[4] Mazlack, L.J., He, A., Zhu, Y., Coppock, S. “A rough set approach in choosing partitioning attributes”. Proceedings of the ISCA 13th, International Conference, CAINE-2000, 2000, 1–6.
[5] Parmar, D., Wu, T. and Blackhurst, J. ”MMR: An algorithm for clustering categorical data using rough set theory”. Data and Knowledge Engineering 63, 2007, 879–893.
[6] Pawlak, Z. and Skowron, A. “Rudiments of rough sets”. Information Sciences, 177 (1), 2007, 3–27.
[7] Yao, Y.Y. “Two views of the theory of rough sets in finite universes”. Approximate Reasoning, 15 (4), 1996, 191–317.
[8] Hu, X. “Knowledge discovery in databases: An attribute oriented rough set approach”. PhD thesis, University of Regina, 1995.
[9] http://archive.ics.uci.edu/ml/datasets/Soybean+%28Small%29
[10] http://archive.ics.uci.edu/ml/datasets/Zoo

Authors

Tutut Herawan
He is a Ph.D. candidate in Data Mining at Universiti Tun Hussein Onn Malaysia (UTHM). His research area includes Data Mining, KDD and Real Analysis.

Rozaida Ghazali
She received her B.Sc. (Hons) degree in Computer Science from Universiti Sains Malaysia, and M.Sc. degree in Computer Science from Universiti Teknologi Malaysia. She obtained her Ph.D. degree in Higher Order Neural Networks at Liverpool John Moores University, UK. She is currently a teaching staff at Faculty of Information technology and Multimedia, Universiti Tun Hussein Onn Malaysia (UTHM). Her research area includes neural networks, fuzzy logic, financial time series prediction and physical time series forecasting.

Iwan Tri Riyadi Yanto
He is a M.Sc. candidate in Data Mining at Universiti Tun Hussein Onn Malaysia (UTHM). His research area includes Data Mining, KDD and Real Analysis.

Mustafa Mat Deris
He received the B.Sc. from University Putra Malaysia, M.Sc. from University of Bradford, England and Ph.D. from University Putra Malaysia. He is a professor of computer science in the Faculty of Information Technology and Multimedia, UTHM, Malaysia. His research interests include distributed databases, data grid, database performance issues and data mining. He has published more than 80 papers in journals and conference proceedings. He was appointed as one of editorial board members for International Journal of Information Technology, World Enformatika Society, a reviewer of a special issue on International Journal of Parallel and Distributed Databases, Elsevier, 2004, a special issue on International Journal of Cluster Computing, Kluwer, 2004, IEEE conference on Cluster and Grid Computing, held in Chicago, April, 2004, and Malaysian Journal of Computer Science. He has served as a program committee member for numerous international conferences/workshops including Grid and Peer-to-Peer Computing, (GP2P 2005, 2006), Autonomic Distributed Data and Storage Systems Management (ADSM 2005, 2006), WSEAS, International Association of Science and Technology, IASTED on Database, etc.