Kỷ yếu Hội nghị Khoa học Quốc gia lần thứ IX "Nghiên cứu cơ bản và ứng dụng Công nghệ thông tin (FAIR'9)"; Cần Thơ, ngày 4-5/8/2016
DOI: 10.15625/vap.2016.0005

AN INFORMATION-THEORETIC METRIC BASED METHOD FOR SELECTING CLUSTERING ATTRIBUTE

Pham Cong Xuyen, Đo Si Truong, Nguyen Thanh Tung
Lac Hong University
pcxuyen@lhu.edu.vn, truongds@lhu.edu.vn, nttung@lhu.edu.vn

ABSTRACT*—Clustering problem appears in many different fields like Data Mining, Pattern Recognition, Bioinfor-matics, etc. The basic objective of clustering is to group objects into clusters so that objects in the same cluster are more similar to one another than they are to objects in other clusters. Recently, many researchers have contributed to categorical data clustering, where data objects are made up of non-numerical attributes. Especially, rough set theory based attribute selection clustering approaches for categorical data have attracted much attention. The key to these approaches is how to select only one attribute that is the best to cluster the objects at each time from many candidates of attributes.*

In this paper, we review three rough set based techniques: Total Roughness (TR), Min-Min Roughness (MMR) and Maximum Dependency Attribute (MDA), and propose MAMD (Minimum value of Average Mantaras Distance), an alternative algorithm for hierarchical clustering attribute selection. MAMD uses Mantaras metric which is an information-theoretic metric on the set of partitions of a finite set of objects and seeks to determine a clustering attribute such that the average distance between the partition generated by this attribute and the partitions generated by other attributes of the objects has a minimum value. To evaluate and compare MAMD with three rough set based techniques, we use the concept of average intra-class similarity to measure the clustering quality of selected attribute. The experiment results show that the clustering quality of the attribute selected by our method is higher than that of attributes selected by TR, MMR and MDA methods.

Keywords*— Data Mining, Hierarchical clustering, Categorical data, Rough sets, Clustering attribute selection.*

I. INTRODUCTION

During the last two decades, data mining has emerged as a rapidly growing interdisciplinary field which merges together databases, statistics, machine learning and related areas in order to extract useful knowledge from data (Han and Kamber, 2006).

Clustering is one of fundamental operations in data mining. It can be defined as as follows. Let $D = \{x_1, x_2, \dots, x_n\}$ be the set of n objects, where each x_i is an N dimensional vector in the given feature space. The clustering activity is to find clusters/groups of objects in such a way that objects within the same cluster have a high degree of similarity, while objects belonging to different clusters have a high degree of dissimilarity [6].

Clustering problem appears in many different domains such as pattern recognition, computer vision, biology, medicine, information retrieval, etc. At present, there exist a large number of clustering algorithms in the literature. Types of clustering are divided broadly into hierarchical and non-hierarchical clustering. Non-hierarchical clustering methods create a single partition of the dataset optimizing a criterion function. Hierarchical clustering methods create a sequence of nested partitions of the dataset.

Most of the earlier works on clustering has been focused on numerical data whose inherent geometric properties can be exploited to naturally define distance functions between data points. However, data mining applications frequently involve many datasets that also consist of categorical attributes on which distance functions are not naturally defined. Recently, clustering categorical data have attracted much attention from the data mining research community [1, 4, 7, 8, 11, 12, 14]. One of the techniques of categorical data clustering was implemented by introducing a series of clustering attributes, in which one of the attributes is selected and used to divide the objects at each time until all objects are clustered. To this, one practical problem is faced: for many candidates of attributes, we need to select only one at each time that is the best attribute to cluster the objects according to some predefined criterion.

Recently, there has been works in the area of applying rough set theory to handle uncertainty in the process of selecting clustering attributes [7, 9, 11, 12]. Mazlack et al. [11] proposed a technique using the average of the accuracy of approximation in the rough set theory called total roughness (TR), where the higher the total roughness is, the higher the accuracy of selecting clustering attribute. Parmar et al. [12] proposed the MMR (Min–Min–Roughness) algorithm, which is a ''purity'' rough set-based hierarchical clustering algorithm for categorical data. The MMR algorithm determines the clustering attribute by MR (Min–Roughness) criterion. However, as Herawan et al. has proven in [7], MMR is the complementary of TR and with this technique, the complexity is still an issue due to all attributes are considered to obtain the clustering attribute. In order to solve these problems, Herawan et al. [7] proposed a new technique called maximum dependency attributes (MDA), which is based on rough set theory by taking into

account the dependency of attributes of the database. According to Herawan et al. [7], MDA technique provides better performance than TR and MMR. However, there is an inherent similarity among TR, MMR and MDA, although they look different. The similarity lies that the values of the three techniques are all mainly determined by the cardinality of lower approximation of an attribute with respect to other attributes.

In this paper, we review three rough set based techniques: Total Roughness (TR), Min-Min Roughness (MMR) and Maximum Dependency Attribute (MDA), and propose MAMD (Minimum value of Average Mantaras Distance), an alternative algorithm for hierarchical clustering attribute selection. MAMD uses Mantaras metric which is an information-theoretic metric on the set of partitions of a finite set of objects and seeks to determine a clustering attribute such that the average distance between the partition generated by this attribute and the partitions generated by other attributes of the objects has a minimum value. To evaluate and compare MAMD with three rough set based techniques, we use the concept of average intra-class similarity to measure the clustering quality of selected attribute. The experiment results show that the clustering quality of the attribute selected by our method is higher than that of attributes selected by TR, MMR and MDA methods.

II. PRELIMINARIES

In this section, some basic notions are briefly reviewed. In Section 2.1, we provide the basic concepts of rough set theory [13] and in Section 2.2 we describe Mantaras metric on the set of partitions of a finite set [10].

A. Rough Set Theory

An information system is a quadruple tuple $S = (U, A, V, f)$, where U is a non-empty finite set of objects, A is a nonempty finite set of attributes, $V = \bigcup_{a \in A} V_a$ where V_a is a set of all values of attribute a, and $f: U \times A \to V$ is a function, called information function, that assigns value a $f(u, a) \in V_a$ for every $(u, a) \in U \times A$.

Definition 1. Let $S = (U, A, V, f)$ be an information system, $B \subseteq A$. Two elements $x, y \in U$ is said to be B-indiscernible in S if and only if $f(x, a) = f(y, a)$, for every $a \in B$.

We denote the indiscernibility relation induced by the set of attributes B by $IND(B)$. Obviously, $IND(B)$ is an equivalence relation and it induces unique partition (clustering) of U. The partition of U induced by $IND(B)$ in $S = (U, A, V, f)$ denoted by P_B and the equivalence class in the partition P_B containing $x \in U$, denoted by $[x]_B$.

Definition 2. Let $S = (U, A, V, f)$ be an information system, $B \subseteq A$ and $X \subseteq U$. 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

$$\underline{B}(X) = \{x \in U | [x]_B \subseteq X\} \text{ and } \overline{B}(X) = \{x \in U | [x]_B \cap X \neq \emptyset\} \quad (1)$$

These definitions state that object $x \in \underline{B}X$ certainly belongs to X, whereas object $x \in \overline{B}X$ could belong to X. Obviously, there is $\underline{B}X \subseteq X \subseteq \overline{B}X$ and X is said to be definable if $\underline{B}X = \overline{B}X$. Otherwise, X is said to be rough with B-boundary $BN_B(X) = \overline{B}X - \underline{B}X$.

Definition 3. Let $S = (U, A, V, f)$ be an information system, $B \subseteq A$ and $X \subseteq U$. The accuracy of approximation of X with respect to B is defined as:

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

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

Obviously, $0 \leq \alpha_B(X) \leq 1$. If $\alpha_B(X) = 1$, then $\underline{B}X = \overline{B}X$. The B-boundary of X is empty, and X is crisp with respect to B. If $\alpha_B(X) < 1$, then $\underline{B}X \subset \overline{B}X$. The B-boundary of X is not empty, and X is rough with respect to B.

Definition 4. Let $S = (U, A, V, f)$ be an information system, $B \subseteq A$ and $X \subseteq U$. The roughness of X with respect to B is defined as:

$$\rho_B(X) = 1 - \frac{|\underline{B}(X)|}{|\overline{B}(X)|} \quad (3)$$

Definition 5. Let $S = (U, A, V, f)$ be an information system. For $P, Q \subseteq A$, it is said that Q depends on P in a degree k $(0 \leq k \leq 1)$, denoted by $P \Rightarrow_k Q$, if

$$k = \frac{\sum_{X \in Q} |\underline{P}(X)|}{|U|} \quad (4)$$

B. Mantaras metric

Definition 4. Let $S = (U, A, V, f)$ be an information system, $X \subseteq A$ and $P_X = \{X_1, X_2, \dots, X_m\}$. The entropy of partition P_X is defined as:

$$E(P_X) = -\sum_{i=1}^{m} P(X_i)\log_2 P(X_i) \tag{5}$$

where $P(X_i) = |X_i|/|U|$, and we define $0\log_2 0 = 0$.

Definition 5. Let $S = (U, A, V, f)$ be an information system, $X, Y \subseteq A$, $P_X = \{X_1, X_2, \dots, X_m\}$, and $P_Y = \{Y_1, Y_2, \dots, Y_n\}$. The conditional entropy of partition P_X with respect to partition P_Y is defined as:

$$E(P_X|P_Y) = -\sum_{j=1}^{n} P(Y_j)\sum_{i=1}^{m} P(X_i|Y_j)\log_2 P(X_i|Y_j) \tag{6}$$

where $P(X_i|Y_j) = |X_i \cap Y_j|/|Y_j|$, $i = 1,2, \dots, m$ and $j = 1,2, \dots, n$.

Definition 6. Let $S = (U, A, V, f)$ be an information system, $X, Y \subseteq A$, $P_X = \{X_1, X_2, \dots, X_m\}$, and $P_Y = \{Y_1, Y_2, \dots, Y_n\}$. The joint entropy of partitions P_X and P_Y is defined as:

$$E(P_X, P_Y) = -\sum_{i=1}^{m}\sum_{j=1}^{n} P(X_i, Y_j)\log_2 P(X_i, Y_j) \tag{7}$$

where $P(X_i, Y_j) = |X_i \cap Y_j|/|U|$, $i = 1,2, \dots, m$ and $j = 1,2, \dots, n$.

From formulas (4) (5) and (6) we have:

$$E(P_X|P_Y) = E(P_X, P_Y) - E(P_Y) \tag{8}$$

Proposition 1. [10**]** The measure

$$d(P_X, P_Y) = E(P_X|P_Y) + E(P_Y|P_X) \tag{9}$$

is a metric on the set of partitions of U, that is, for any partitions P_X, P_Y, and P_Z on U it satisfies

(i) $d(P_X, P_Y) \geq 0$ and the equality holds iff $P_X = P_Y$

(ii) $d(P_X, P_Y) = d(P_Y, P_X)$

(iii) $d(P_X, P_Y) + d(P_Y, P_Z) \geq d(P_X, P_Z)$.

Note that, from formula (8), we can write:

$$d(P_X, P_Y) = 2E(P_X, P_Y) - E(P_X) - E(P_Y) \tag{10}$$

III. THREE ROUGH SET-BASED TECHNIQUES

Let $S = (U, A, V, f)$ be an information system , $a_i \in A$, $V(a_i)$ refers to the set of values of attribute a_i, $X(a_i = \alpha)$ is a subset of objects having one specific value, α, of attribute a_i, that is, $X(a_i = \alpha)$ is a class of objects induced by indiscernibility relation $IND(a_i)$, $\underline{X_{a_j}(a_i = \alpha)}$ refers to the lower approximation, and $\overline{X_{a_j}(a_i = \alpha)}$ refers to the upper approximation with respect to a_j.

A. TR (***Total Roughness***) *Technique* **[11]**

Input: Dataset (information system) S without clustering attribute

Output: Clustering attribute

Begin

Step 1: Compute the equivalence classes using the indiscernibility relation on each attribute.

Step 2: For each a_i determine its **mean roughness** $Rough_{a_j}(a_i)$ with respect to all a_j, $j \neq i$, by the following formula

$$Rough_{a_j}(a_i) = \frac{\sum_{k=1}^{|V(a_i)|} R_{a_j}(X \mid a_i = \alpha_k)}{|V(a_i)|} \tag{11}$$

where

$$R_{a_j}(X \mid a_i = \alpha) = \frac{\left|\underline{X_{a_j}(a_i = \alpha)}\right|}{\left|\overline{X_{a_j}(a_i = \alpha)}\right|} \tag{12}$$

Step 3: For each $a_i \in A$ compute its **total roughness** with respect to a_j, $i \neq j$, by the following formula

$$TR(a_i) = \frac{\sum_{(j=1)\wedge(i\neq j)}^{|A|} Rough_{a_j}(a_i)}{|A| - 1} \tag{13}$$

Step 4. Select the attribute a_i^* with the maximum value of *TR* as clustering attribute, i.e.

$$a_i^* = \text{argmax}_{a_j \in A}\{TR(a_j)\} \tag{14}$$

End

B. MMR (*Min–Min-Roughness*) *Technique* [12]

Input: Dataset (information system) S without clustering attribute

Output: Clustering attribute

Begin

Step 1: Compute the equivalence classes using the indiscernibility relation on each attribute.

Step 2: For each a_i determine its **mean roughness** $Rough_{a_j}(a_i)$ with respect to all a_j, $j \neq i$, by the following formula

$$Rough_{a_j}(a_i) = \frac{\sum_{k=1}^{|V(a_i)|} R_{a_j}(X \mid a_i = \alpha_k)}{|V(a_i)|} \tag{15}$$

where

$$R_{a_j}(X \mid a_i = \alpha) = 1 - \frac{\left|\underline{X_{a_j}(a_i = \alpha)}\right|}{\left|\overline{X_{a_j}(a_i = \alpha)}\right|} \tag{16}$$

Step 3: For each a_i determine its **minimum roughness** $MR(a_i)$ by the following formula

$$MR(a_i) = \min_{(a_j \in A)\wedge(j\neq i)} \left(Rough_{a_j}(a_i)\right) \tag{17}$$

Step 4. Select the attribute a_i^* with the **minimum** value of *MR* as clustering attribute, i.e.

$$a_i^* = \text{argmin}_{a_j \in A}\{MR(a_j)\} \tag{18}$$

End

C. MDA (*Maximum degree of Dependency of Attributes*) *Technique* [7]

Input: Dataset (information system) S without clustering attribute

Output: Clustering attribute

Begin

Step 1. Compute the equivalence classes using the indiscernibility relation on each attribute.

Step 2. For each a_i determine the dependency degree of attribute a_i with respect to all a_j, where $j \neq i$. by the following formula

$$\gamma_{a_j}(a_i) = \frac{\sum_{X \in U/a_i} \left|\underline{a_j} X\right|}{|U|} \tag{19}$$

Step 3. Select the **maximum of dependency degree** $MD(a_i)$ of each attribute a_i ($a_i \in A$) as following

$$MD(a_i) = \max_{(a_j \in A) \wedge (i \neq j)} \left(\gamma_{a_j}(a_i) \right) \quad (20)$$

Step 4. Select the attribute a_i^* with the **maximum** value of *MD* as clustering attribute, i.e.

$$a_i^* = \text{argmax}_{a_j \in A} \{MD(a_j)\} \quad (21)$$

End

IV. MAMD (MINIMUM AVERAGE MANTARAS DISTANCE) TECHNIQUE

In this section we present MAMD technique, which is based on Minimum Average Mantaras Distance, to select clustering attribute.

Definition 7. Let $S = (U, A, V, f)$ be an information system, and $a_i \in A$ is an attribute, the Average Mantaras Distance of a_i to all $a_j, j \neq i$, is defined by the following formula:

$$AMD(a_i) = \frac{\sum_{j=1, j \neq i}^{|A|} d\left(P_{a_i}, P_{a_j}\right)}{|A| - 1} \quad (22)$$

where P_{a_i}, P_{a_j} are the partitions of U induced by $IND(a_i)$ and $IND(a_j)$ respectively.

In the above definitions, $d\left(P_{a_i}, P_{a_j}\right)$ is a measurement to the distance between P_{a_i} and P_{a_j}. From the view of clustering, the lower AMD is, the higher the crispness of the clustering. Based on the above definition, we present the MAMD algorithm as follows.

Input: Dataset (information system) S without clustering attribute

Output: Clustering attribute

Begin

Step 1. For each attribute $a_i \in A$, compute the equivalence classes of partition P_{a_i} induced by indiscernibility relation $IND(a_i)$.

Step 2. For each attribute $a_i \in A$ compute the condition entropy of partition P_{a_i} with respect to partition P_{a_j}, where $i \neq j$:

$$E(P_X|P_Y) = -\sum_{j=1}^{n} P(Y_j) \sum_{i=1}^{m} P(X_i|Y_j) \log_2 P(X_i|Y_j) \quad (23)$$

where $P_{a_i} = \{X_1, X_2, \ldots, X_m\}$, $P_{a_j} = \{Y_1, Y, \ldots, Y_n\}$, and $P(X_i|Y_j) = |X_i \cap Y_j| / |Y_j|$, $i = 1,2, \ldots, m$ and $j = 1,2, \ldots, n$.

Step 3. For each every pair $a_i, a_j \in A$, where $i \neq j$, compute the distance between two partitions P_{a_i} and P_{a_j} using Mantaras metric:

$$d\left(P_{a_i}, P_{a_j}\right) = E\left(P_{a_i} \middle| P_{a_j}\right) + E\left(P_{a_j} \middle| P_{a_i}\right) \quad (24)$$

Step 4. For each attribute $a_i \in A$, compute the Average Mantaras Distance $AMD(a_i)$ according to (22).

Step 5. Select the attribute with the lowest AMD as clustering attribute.

End

Let us illustrate the MAMD algorithm by an example.

Example. Table 1 shows Credit dataset as in **[7]**. There are ten objects with five categorical attributes: Magazine Promotion (MP), Watch Promotion (WP), Life Insurance Promotion (LIP), Credit Card Insurance (CCI), and Sex (S).

First, we deal with attribute MP. The partition of U induced by attribute MP is:

P_{MP} ={{1, 2, 4, 5, 7, 9, 10}, {3, 6, 8}}.

We have:

$$E(P_{\text{MP}}) = -\frac{7}{10} log_2 \frac{7}{10} - \frac{3}{10} log_2 \frac{3}{10} = 0.881 ;$$

The partition of U induced by attribute WP is:

$P_{\mathrm{WP}} = \{\{1, 3, 5, 6, 7, 9\}, \{2, 4, 8, 10\}\}$

$$E(P_{\mathrm{WP}}) = -\frac{6}{10} log_2 \frac{6}{10} - \frac{4}{10} log_2 \frac{4}{10} = 0.971;$$

Table 1. The Credit dataset

#	MP	WP	LIP	CCI	S
1	yes	no	no	no	Male
2	yes	yes	yes	no	Female
3	no	no	no	no	Male
4	yes	yes	yes	yes	Male
5	yes	no	yes	no	Female
6	no	no	no	no	Female
7	yes	no	yes	yes	Male
8	no	yes	no	no	Male
9	yes	no	no	no	Male
10	yes	yes	yes	no	Female

The partition of U induced by {MP,WP} is:

$P_{\{\mathrm{MP,WP}\}} = \{\{1, 5, 7, 9\}, \{2, 4, 10\}, \{3, 6\}, \{8\}\}$

$$E\left(P_{\{\mathrm{MP,WP}\}}\right) = -\frac{4}{10} log_2 \frac{4}{10} - \frac{3}{10} log_2 \frac{3}{10} - \frac{2}{10} log_2 \frac{2}{10} - \frac{1}{10} log_2 \frac{1}{10} = 1.846;$$

Applying Eq. (10), we get the distance between P_{MP} and P_{WP} as follows:

$$d(P_{\mathrm{MP}}, P_{\mathrm{WP}}) = 2E\left(P_{\{\mathrm{MP,WP}\}}\right) - E(P_{\mathrm{MP}}) - E(P_{\mathrm{WP}}) = 1.840;$$

With the same process, we can get the distance between MP and other attributes:

$$d(P_{\mathrm{MP}}, P_{\mathrm{LIP}}) = 1.090,\ d(P_{\mathrm{MP}}, P_{\mathrm{CCI}}) = 1.368,\ d(P_{\mathrm{MP}}, P_{\mathrm{S}}) = 1.840,$$

The Average Mantaras Distance AMD of attribute MP can be computed by Eq.(22) as:

$$AMD(\mathrm{MP}) = \frac{1.840 + 1.090 + 1.368 + 1.840}{4} = 1.535$$

With the same process as MP, we can deal with other attributes . The AMD and $MAMD$ of all attributes are summarized in Table 2.

Table 2. All values of distance and average distance between attributes in Table 1

Attribute	Distance (w.r.t.…)					AMD
	MP	WP	LIP	CCI	S	
MP	0.000	1.840	1.090	1.368	1.840	1.535
WP	1.841	0.000	1.722	1.678	1.902	1.786
LIP	1.090	1.722	0.000	1.249	1.722	1.446
CCI	1.368	1.678	1.249	0.000	1.351	**1.412**
S	1.840	1.902	1.722	1.351	0.000	1.704

From Table 2, we can see that attribute CCI has the smallest AMD; therefore CCI is selected as clustering attribute using MAMD algorithm.

V. COMPARISON TESTS

A. clustering quality measure

The four techniques TR, MMR, MDA and MAMD techniques use different methods for selecting clustering attribute. Measuring the clustering quality of selected attribute in a just manner is a non-trivial task. Since the goal of cluster analysis is to group data with similar characteristics, we use average intra-class similarity to measure the quality.

Definition 8. Let $S = (U, A, V, f)$ be an information system and suppose that all attributes in A are categorical. Then the similarity between two objects x_i and x_j in U is defined as:

$$s(x_i, x_j) = \frac{\left|\left\{\{a_k \in A\} \middle| f(x_i, a_k) = f(x_j, a_k)\right\}\right|}{|A|} \quad (25)$$

Definition 9. Let $S = (U, A, V, f)$ be an information system. Suppose $a_j \in A$ is selected as clustering attribute and the clustering (partition) induced by a_j is $P_{a_j} = \{X_1, X_2, \dots, X_m\}$ where $X_i = \{x_{i1}, x_{i2}, \dots, x_{i|X_i|}\}$, $i = 1, 2, \dots, m$. The average similarity (AS) of object x_{ij} with respect to other objects in X_i is defined as:

$$AS(x_{ij}) = \frac{\sum_{k=1, k\neq j}^{|X_i|} s(x_{ij}, x_{ik})}{|X_i| - 1} \quad (26)$$

Definition 10. Let $S = (U, A, V, f)$ be an information system. Suppose $a_j \in A$ is selected as clustering attribute and the clustering (partition) induced by a_j is $P_{a_j} = \{X_1, X_2, \dots, X_m\}$ where $X_i = \{x_{i1}, x_{i2}, \dots, x_{i|X_i|}\}$, $i = 1, 2, \dots, m$. The intra-class similarity (CS) of cluster X_i is defined as:

$$CS(X_i) = \frac{\sum_{j=1}^{|X_i|} AS(x_{ij})}{|X_i|} \quad (27)$$

Definition 11. Let $S = (U, A, V, f)$ be an information system. Suppose $a_j \in A$ is selected as clustering attribute and the clustering (partition) induced by a_j is $P_{a_j} = \{X_1, X_2, \dots, X_m\}$ where $X_i = \{x_{i1}, x_{i2}, \dots, x_{i|X_i|}\}$, $i = 1, 2, \dots, m$. The average intra-class similarity (ACS) of clustering induced by a_j is defined as:

$$ACS(a_j) = \frac{\sum_{i=1}^{m} CS(X_i)}{m}. \quad (28)$$

The higher the average intra-class similarity is the higher the clustering quality of the selected attribute.

B. Datasets for testing and results

The data sets of four test cases, as in [7], are presented in Table 1, Table 3, Table 4 and Table 5. These are Credit, Animal world, Parmar, and student's enrollment qualification data sets.

Table 3. The animal world data set

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	craw	grain	N	Y	Y
Eagle	N	N	forward	Y	craw	meat	N	Y	N
Viper	N	pointed	forward	N	N	meat	N	N	N

Table 4. The Parmar data set

Rows	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

Table 5. The student's enrollment qualification data set

Student	Degree	English	Experience	IT	Maths	Programming	Statistics
1	PhD	Good	Medium	Good	Good	Good	Good
2	PhD	Medium	Medium	Good	Good	Good	Good
3	M.Sc	Medium	Medium	Medium	Good	Good	Good
4	M.Sc	Medium	Medium	Medium	Good	Good	Medium
5	M.Sc	Medium	Medium	Medium	Medium	Medium	Medium
6	M.Sc	Medium	Medium	Medium	Medium	Medium	Medium
7	B.Sc	Medium	Good	Good	Medium	Medium	Medium
8	B.Sc	Bad	Good	Good	Medium	Medium	Good

We have installed all the four techniques TR, MMR, MDA and MAMD, in R programming language using RoughSets package.

As the results of computations, Table 6 shows the clustering attributes chosen by the techniques in each data set.

Table 6. Clustering attributes selected by the techniques in data set

	Datasets			
	Credit	Animal	Parmar	Student
TR	CCI	Hair	Var1	Experience
MMR	CCI	Hair	Var1	Experience
MDA	CCI	Hair	Var1	Experience
MAMD	CCI	Teeth	Var2	Degree

From Table 6, we can see that for all the four considered data sets, three techniques TR, MMR and MDA choose same attribute as the clustering attribute. Our technique MAMD chooses same attribute CCI in Credit data set, but other attributes in Animal world, Parmar and Student's enrollment qualification data sets.

Now, let us measure the clustering quality of attributes chosen by TR, MMR, MDA and MAMD in these three data sets.

We take attribute Hair in Credit data set as an example to calculate the average intra-class similarity.

The partition of U induced by attribute Hair consists of two equivalence classes:

$X_1 = X(\text{Hair} = \text{Y}) = \{\text{Tiger}, \text{Cheetah}, \text{Giraffe}, \text{Zebra}\}$,

$X_2 = X(\text{Hair} = \text{N}) = \{\text{Ostrich}, \text{Penguin}, \text{Albatross}, \text{Eagle}, \text{Viper}\}$.

We take animal Tiger in X_1 as an example to calculate the similarity, average similarity. Applying Eq. (25), we have

$S(\text{Tiger}, \text{Cheetah}) = 1$, $S(\text{Tiger}, \text{Giraffe}) = 0.444$, $S(\text{Tiger}, \text{Zebra}) = 0.444$.

Applying Eq. (26), the average similarity of Tiger with respect to other animals in X_1 is calculated as follows:

$$AS(\text{Tiger}) = \frac{1 + 0.444 + 0.444}{3} = 0.630$$

With the same process, the similarity and average similarity of other animals in X_1 are calculated and summarized as in Table 7.

Table 7. The similarity, AS and CS of all animals in X_1 induced by Hair

Animal	Tiger	Cheetah	Giraffe	Zebra	AS	CS
Tiger	-	1.000	0.444	0.444	0.630	0.630
Cheetah	1.000	-	0.444	0.444	0.630	
Giraffe	0.444	0.444	-	1.000	0.630	
Zebra	0.444	0.444	1.000	-	0.630	

Applying Eq. (27), the intra-class similarity of X_1 is calculated below.

$$CS(X_1) = \frac{0.630 + 0.630 + 0.630 + 0.630}{4} = 0.630$$

Using the same way, we obtain $CS(X_2) = 0.544$.

Lastly, using Eq. (28), the average intra-class similarity is calculated as follows:

$$ACS(\text{Hair}) = \frac{0.630 + 0.544}{2} = 0.587.$$

With the same process, we have computed the average intra-class similarity of clustering induced by Teeth in "Animal world", by Var1 and by Var2 in Parmar, by Experience and by Degree in Student dataset. The computation results are given in Table 8.

Table 8. The average intra-class similarity (ACS) of clustering induced by attributes in the datasets

	Selected attributes and their ACS values		
	Animal	**Parmar**	**Student**
TR	Hair **0.587**	Var1 **0.536**	Experience **0.638**
MMR	Hair **0.587**	Var1 **0.536**	Experience **0.638**
MDA	Hair **0.587**	Var1 **0.536**	Experience **0.638**
MAMD	Teeth **0.784**	Var2 **0.555**	Degree **0.770**

The results in Table 8 show that the clustering quality of attribute selected using MAMD technique is higher than that of attribute selected by TR, MMR and MDA techniques.

VI. CONCLUSION

In the recent years, some techniques applying rough set theory for selecting clustering attributes have been proposed. However, although they look different, there is an inherent similarity among them, and the computational complexity is still an issue.

In this paper, we review three rough set based techniques: Total Roughness (TR), Min-Min Roughness (MMR) and Maximum Dependency Attribute (MDA), and propose MAMD (Minimum value of Average Mantaras Distance), an alternative algorithm for hierarchical clustering attribute selection. MAMD uses Mantaras metric which is an information-theoretic metric on the set of partitions of a finite set of objects and seeks to determine a clustering attribute such that the average distance between the partition generated by this attribute and the partitions generated by other attributes of the objects has a minimum value. To evaluate and compare MAMD with three rough set based techniques, we use the concept of average intra-class similarity to measure the clustering quality of selected attribute. The experiment results show that the clustering quality of the attribute selected by our method is higher than that of attributes selected by TR, MMR and MDA methods. The proposed approach could be integrated into clustering algorithm based on attributes selection for categorical data.

REFERENCES

[1] Barbara, D., Li, Y., Couto, J.: COOLCAT: an entropy-based algorithm for categorical clustering. In: *Proc. of CIKM 2002*, 582–589, 2002.

[2] Cao F. Y., Liang J. Y. , Li D. Y., Bai L., A new initialization method for categorical data clustering, *Expert Syst. Appl*. 36, 10223–10228, 2009.

[3] Ganti, V., Gehrke, J., Ramakrishnan, R.: CACTUS – clustering categorical data using summaries. In: *Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 73–83, 1999.

[4] Guha, S., Rastogi, R., Shim, K.: ROCK: a robust clustering algorithm for categorical attributes. *Information Systems*, 25(5), 345–366, 2000.

[5] Huang, Z.: Extensions to the k-averages algorithm for clustering large datasets with categorical values. *Data Mining and Knowledge Discovery*, 2(3), 283–304, 1998.

[6] Han J., and Kamber M., *Data Mining: Concepts and Techniques*, Morgan Kanufmann, 2006.

[7] Herawan, T., Deris, M. M., Abawajy, J. H.: A rough set approach for selecting clustering attribute. *Knowledge-Based Systems* 23, 220–231, 2010.

[8] Jain A. K., Data clustering: 50 years beyond k-averages, Pattern Recogn. Lett., 31(8), 651–666, 2010.

[9] Jyoti Dr., Clustering categorical data using rough sets: a review. *International Journal of Advanced Research in IT and Engineering*, Vol. 2, No. 12, December, 30-37, 2013.

[10] Lopez de Mantaras R., A Distance-Based Attribute Selection Measure for Decision Tree Induction. *Machine Learning*, 6, N^o 1, 81-92, 1991.

[11] 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), 1–6, 2000.

[12] 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.

[13] Pawlak, Z.: Rough sets. *International Journal of Computer and Information Science*, 11, 341–356, 1982.

[14] Suchita S. Mesakar, M. S. Chaudhari, Review Paper On Data Clustering Of Categorical Data. *International Journal of Engineering Research & Technology*, Vol. 1 Issue 10, December, 2012.

MỘT PHƯƠNG PHÁP LỰA CHỌN THUỘC TÍNH GOM CỤM SỬ DỤNG METRIC LÝ THUYẾT THÔNG TIN

Phạm Công Xuyên, Đỗ Sĩ Trường, Nguyễn Thanh Tùng

TÓM TẮT *— Bài toán gom cụm dữ liệu xuất hiện trong nhiều lĩnh vực khác nhau như Khai thác dữ liệu, Nhận dạng, Tin-sinh học, vv. Mục tiêu cơ bản của gom cụm là nhóm đối tượng thành các cụm sao cho các đối tượng trong cùng một cụm thì tương tự như nhau hơn là các đối tượng từ các cụm khác nhau. Gần đây, nhiều nhà nghiên cứu quan tâm đến vấn đề gom cụm dữ liệu phạm trù (categorical), trong đó các đối tượng dữ liệu được mô tả bởi các thuộc tính không phải thuộc tính số. Đặc biệt, phương pháp tiếp cận sử dụng lý thuyết tập thô trong gom cụm phân cấp (hierarchical) dữ liệu phạm trù đã thu hút nhiều sự chú ý. Chìa khóa của các phương pháp này là làm thế nào để chọn được một thuộc gom cụm tốt nhất tại mỗi thời điểm trong số nhiều thuộc tính ứng viên.*

Trong bài báo này, chúng tôi xem xét ba kỹ thuật dựa trên lý thuyết tập thô: Total Roughness (TR), Min-Min Roughness (MMR) và Maximum Dependency Attribute (MDA), và đề xuất MAMD (Minimum value of Average Mantaras Distance), một thuật toán mới cho việc lựa chọn thuộc tính phân cụm theo tiếp cận phân cấp. MAMD sử dụng metric Mantaras, một metric lý thuyết thông tin trên tập các phân hoạch của một tập hợp gồm hữu hạn các đối tượng và tìm cách xác định thuộc tính gom cụm sao cho khoảng cách trung bình giữa phân hoạch sinh ra bởi thuộc tính này và các phân hoạch sinh ra bởi các thuộc tính khác đạt giá trị nhỏ nhất. Để đánh giá và so sánh MAMD với ba kỹ thuật dựa trên lý thuyết tập thô, chúng tôi sử dụng khái niệm "Độ tương tự trung bình bên trong cụm" của một phép gom cụm để đo lường chất lượng gom cụm của thuộc tính được chọn. Kết quả thực nghiệm cho thấy chất lượng gom cụm của thuộc tính chọn được bằng phương pháp của chúng tôi là cao hơn so với các thuộc tính chọn bởi các phương pháp TR, MMR và MDA. Do đó, MAMD có thể được sử dụng như là một kỹ thuật hiệu quả lựa chọn thuộc tính trong phân cụm phân cấp.

Từ khóa *— Gom cụm, Lý thuyết tập thô, Lựa chọn thuộc tính, Mantaras metric, Gom cụm phân cấp.*