International Journal of Database Theory and Application
Vol.6, No.5 (2013), pp.53-66
http://dx.doi.org/10.14257/ijdta.2013.6.5.06

An Algorithm for Selecting Clustering Attribute using Significance of Attributes

W. A. Hassanein[1] and Amr A. Elmelegy[*1]
[1]*Mathematics Department*
Faculty of science, Tanta University, Tanta, Egypt
wfanwar@yahoo.com,
amrelmelegy294@yahoo.com[*] (corresponding author).

Abstract

There are fewer techniques to group objects having similar characteristics deal with categorical data ,but some are of them be complicated in the clustering process while others have stability issues. In this paper we represent a new technique which it be more easier than the other techniques in computing the selecting clustering attribute process and at the same time having stability issues besides taking care of handling uncertainty and categorical data together, we called it (maximum significance of attributes) MSA. The proposed technique based on rough set theory by taking into account the concept of significance of attributes of the database. We analyzing and comparing the performance of MSA technique with (bi-clustering) BC, (total roughness) TR, (minimum-minimum roughness) MMR and (maximum dependency of attribute) MDA techniques.

Keywords: *Clustering, Rough Set theory, Categorical data, Significance of attributes, performance*

1. Introduction

A cluster is a collection of data objects that are similar to one another within the same cluster. And are dissimilar to the objects in other clusters. And its many uses, for example manufacturing,medicine, nuclear science, radar scanning and research and development planning. For example,Wu *et al.*, [1] developed a clustering algorithm specifically designed for handling the complexity of gene data. Jiang *et al.*, [2] analyze a variety of cluster techniques, which can be applied for gene expression data. Wong *et al.*, [3] presented an approach used to segment tissues in a nuclear medical imaging method known as positron emission tomography (PET). Haimov *et al.*,[4] used cluster analysis to segment radar signals in scanning land and marine objects. Finally Mathieu and Gibson [5] used the cluster analysis as a part of a decision support tool for large scale research and development planning to identify programs to participate in and to determineresource allocation.

The problem with all the above mentioned algorithms is that they mostly deal with numerical data sets that are those databases having attributes with numeric domains. The basic reason for dealing with numerical attributes is that these are very easy to handle and also it is easy to define similarity on them. Unlike numerical data, categorical data have multi-valued attributes .This, similarity can be defined as common objects, common values for the attributes and the association between two.

In such cases, a number of algorithms for clustering categorical data have been proposed including work by Huang [6], Gibson *et al.*, [7], Guha *et al.*, [8], Ganti *et al.*, [9], and Dempster *et al.*, [10]. While these methods make important contributions to the issue of clustering categorical data, they are not designed to handle uncertainty in the clustering process. This is an important issue in many real world applications where there is often no sharp boundary between clusters.

ISSN: 2005-4270 IJDTA

Recently, there has been work in the area of applying rough set theory to handle uncertainty in the process of selecting clustering attribute, proposed by Z. Pawlak in 1982 [11]. It has been applied to machine learning, intelligent systems, inductive reasoning, pattern recognition, expert systems, data analysis, data mining and knowledge discovery. Rough set theory overlaps with many other theories. Despite this overlap, rough set theory may be considered as an independent discipline in its own right. The main advantage of rough set theory in data analysis is that it does not need any preliminary or additional information about data like probability distributions in statistics, basic probability assignments in Dempster_Shafer theory, a grade of membership or the value of possibility in fuzzy set theory [12].

And in 2010 [Herawan *et al.*,] proposed a technique to selecting clustering attribute: *i.e.*, called maximum dependency of attributes (MDA) It is based on the dependency of attributes using rough set theory in an information system Herawan *et al.*, in 2010 [13].

In this paper, we propose a new technique called maximum significance attribute (MSA) .It is based on the significance of attributes using rough set theory in an information system because we need for another technique in data clustering to improve the clustering process to make it more easier than the other techniques in computing the selecting clustering attribute process and at the same time having stability issues besides taking care of handling uncertainty and categorical data together.

The rest of this paper is organized as follows Section 2 the main concepts important definitions Section 3 we describing the algorithm of MSA technique. Section 4 is the experimental part. Section 5 describes the performance comparison of MSA with BC, TR, MMR and MDA techniques Section 6 describing the conclusions of this work.

2. The Main Concepts of Important Definitions

Definition 1 Information System

In the rough set , information systems are used to represent knowledge, an information system is S=(U,A,V,F) where ; U is a non empty, finite set of objects; A is a non empty, finite set of attributes; $V = \bigcup_{a \in A} V_a$, V_a is the domain (value set) of attribute a; $f : U \times A \to V$ is a total function such that $f(u,a) \in V_a$ for every $(u,a) \in U \times A$,called information (knowledge) function as in[12].

Definition 2 Indiscernibility Relation

(Indiscernibility relation (Ind (B))): Ind (B) is a relation on U see[12]. Given two objects $x_i, x_j \in U$, they are indiscernible by the set of attributes B in A, if and only if $a(x_i) = a(x_j)$ for every $a \in B$. That is, ($x_i, x_j \in Ind(B)$) if and only if $\forall a \in B$ where $B \subseteq A$, $a(x_i) = a(x_j)$.

Definition 3 Equivalence Classes

(Equivalence class ($[x_i]_{Ind(B)}$)) proposed in[12]: Given Ind (B), the set of objects x_i having the same values for the set of attributes in B consists of an equivalences classes, $[x_i]_{Ind(B)}$. It is also known as elementary set with respect to B.

Definition 4 Upper Approximation

Given the set of attributes B in A, set of objects X in U, the lower approximation of X is defined as the union of all the elementary sets which are contained in X. That is $\overline{X}_B = \bigcup\{X_i \mid [X_i]_{Ind(B)} \cap X \neq \phi\}$ in[12].

Definition 5 Lower Approximation

Given the set of attributes B in A, set of objects X in U, the lower approximation of X is defined as in [12] the union of all the elementary sets which are contained in X. That is $\underline{X}_B = \bigcup\{X_i \mid [X_i]_{Ind(B)} \subseteq X\}$

Definition 6 Dependency of Attributes

Suppose S=(U,A,V,F) is information system and let a_i and a_j be any subsets of A. Dependency attribute a_i on a_j in a degree k(0 < k < 1), is denoted by $a_i \Rightarrow a_j$. The degree k Herawan *et al.*, in 2010[13] is defined:

$$K = \gamma_{a_j}(a_i) = \frac{\sum_{X \in U/a_j} \left|\underline{a_i}(X)\right|}{\left|U\right|} \tag{1}$$

Definition 7 Roughness

Suppose that attribute $a_i \in A$ has k-different values, say β_k, k = 1,2,. . .,n. Let $X(a_i = \beta_k)$, k = 1,2,. . .,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,. . .,n, with respect to a_i, where $i \neq j$, denoted by $R_{a_j}(X / a_i = \beta_k)$, as in [14] is defined by

$$R_{a_j}(X / a_i = \beta_k) = \frac{\left|\underline{X_{a_j}(a_i = \beta_k)}\right|}{\left|\overline{X_{a_j}(a_i = \beta_k)}\right|}, \text{K=1,2,…,n} \tag{2}$$

Definition 8 Mean Roughness

From TR technique, the mean roughness of attribute $a_i \in A$ with respect to attribute $a_j \in A$, where $i \neq j$, denoted $Rough_{a_j}(a_i)$, is evaluated as follows

$$Rough_{a_j}(a_i) = \frac{\sum_{k=1}^{|V(a_i)|} R_{a_j}(X / a_i = \beta_k)}{\left|V(a_i)\right|} \tag{3}$$

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

Definition 9 Total Roughness

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

$$TR(a_i) = \frac{\sum_{j=1}^{|A|} Rough_{a_j}(a_i)}{|A|-1} \quad (4)$$

As stated in Mazlack *et al.*, [14] that the highest value of TR, the best selection of clustering attributes.

Definition 10 Minimum-Minimum Roughness

Meanwhile, the value of roughness of MMR technique is the opposite of that TR technique which is equivalent with that has been proposed in [15], *i.e.*

$$MMR_{a_j}(X / a_i = \beta_k) = 1 - R_{a_j}(X / a_i = \beta_k) \quad (5)$$

Where the value of mean roughness of MMR technique is also the opposite of that TR technique
As in [19]

$$MMRough_{a_j}(a_i) = \frac{\sum_{K=1}^{|V(a_i)|} MMR_{a_j}(X / a_i = \beta_k)}{|V(a_i)|}$$

$$= \frac{\sum_{K=1}^{|V(a_i)|}(1 - R_{a_j}(X / a_i = \beta_k))}{|V(a_i)|}$$

$$= \frac{\sum_{K=1}^{|V(a_i)|}1 - \sum_{K=1}^{|V(a_i)|} R_{a_j}(X / a_i = \beta_k)}{|V(a_i)|} = \frac{|V(a_i)|}{|V(a_i)|} - \frac{\sum_{K=1}^{|V(a_i)|} R_{a_j}(X / a_i = \beta_k)}{|V(a_i)|}$$

$$= 1 - Rough_{a_j}(a_i) \quad (6)$$

Definition 11 Significance of Single Attribute

Suppose Significance of single attribute $a_i \in$ A related to $a_j \in A$

$\sigma_{a_j}(a_i) = \gamma_{A'}(a_j) - \gamma_{A''}(a_j)$ Proposed in[18] (7)

Where $A' = A - \{a_j\}$, $A'' = A' - \{a_i\}$

Definition 12 The accuracy

To measure the accuracy of selecting clustering attribute, we use the formula of mean roughness in Eq. (3) to represent all techniques. The higher the mean roughness is the higher the accuracy of the selecting clustering attribute[13].

Definition 13 The Purity Ratio

In order to compare (MSA) with (MMR,TR and MDA) and all other algorithms which have taken initiative to handle categorical data we developed an implementation. The traditional approach for calculating purity of a cluster proposed in [17] is given below.

$$purity = \frac{number\ of\ data\ occuring\ \text{in}\ both\ the\ i^{th} cluster\ and\ its\ corresponding\ class}{the\ number\ of\ data\ in\ the\ set} \quad (8)$$

$$over\ all\ \ purity = \frac{\sum_{i=1}^{\#of\ clusters} purity(i)}{\#of\ clusters} \tag{9}$$

2. Proposed Algorithm

In this section we present our algorithm which we call it MSA. The notations and definitions of concepts have been discussed in the previous section. Suppose that condition attribute set $A = \{a_1, a_2, ..., a_n\}$, the algorithm for solution to single significance of a_i with respect to a_j where $i \neq j$.

Algorithm : MSA
Input : Data set without clustering attribute
Output : Clustering attribute
Begin

1) Get $U / ind(A')$, which A' denotes the family (A) expect $\{a_j\}$ of all equivalence classes of $A - a_j$, written U / A'.
2) Get U / A'', which denotes the equivalence classes of $\{A - \{a_j\}\} - \{a_i\}$ or A'-{ a_i }.
3) Get U / a_j.
4) Get $pos_{A'}(a_j)$.
5) Compute $\gamma_{A'}(a_j)$, which is the dependability of a_j for condition attribute set A'.
6) Get $pos_{A''}(a_j)$.
7) Compute $\gamma_{A''}(a_j)$.
8) Compute $\sigma_{a_j}(a_i)$ as $\sigma_{a_j}(a_i) = \gamma_{A'}(a_j) - \gamma_{A''}(a_j)$.
9) Select the maximum of Significance degree of each attribute.
10) Select the clustering attribute based on the maximum degree of Significance of attributes.

End

Figure 1. The MSA Algorithm

4. Experimental Part

The Case is: The credit card promotion dataset in [16] Table 1 shows the credit card promotion dataset as in [16]. There are five categorical attributes (n = 5): magazine promotion (MP), watch promotion (WP), life insurance promotion (LIP), credit card insurance (CCI) and sex (S). All attributes have two distinct values,(l = 2), *i.e.*, yes and no and ten objects (m = 10) are considered. Notice that with the BC technique, the attribute of the least distinct balanced-valued will be selected as a clustering attribute without consideration of the maximum value of total roughness of each attributes. Thus, attribute LIP will be chosen as a clustering attribute. To illustrate in finding the degree of dependency and significance of attributes, we consider the information system as shown in Table 1

Table 1. A Subset of the Credit Card Promotion Dataset from Acme Credit Card Company Database [16]

Person	Magazine Promotion	Watch Promotion	Life insurance Promotion	Credit card Insurance	Sex
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

4.1. Computational Part

Notice that with the BC technique, the attribute of the least distinct balanced-valued will be selected as a clustering attribute without consideration of the maximum value of total roughness of each attributes. Thus, attribute (LIP) will be chosen as a clustering attribute.

4.1.1. Getting Equivalence Classes

a) $X(MP = yes) = \{1,2,4,5,7,9,10\}$, $X(MP = no) = \{3,6,8\}$,
$U / MP = \{\{1,2,4,5,7,9,10\},\{3,6,8\}\}$
b) $X(WP = yes) = \{2,4,8,10\}$, $X(WP = no) = \{1,3,5,6,7,9\}$,
$U / WP = \{\{2,4,8,10\},\{1,3,5,6,7,9\}\}$
c) $X(LIP = yes) = \{2,4,5,7,10\}$, $X(LIP = no) = \{1,3,6,8,9\}$,
$U / LIP = \{\{2,4,5,7,10\},\{1,3,6,8,9\}\}$
d) $X(CCI = yes) = \{4,7\}$,
$X(CCI = no) = \{1,2,3,5,6,8,9,10\}, U / CCI = \{\{4,7\},\{1,2,3,5,6,8,9,10\}\}$
e) $X(S = yes) = \{1,3,4,7,8,9\}$, $X(S = no) = \{2,5,6,10\}$
$, U / S = \{\{1,3,4,7,8,9\},\{2,5,6,10\}\}$

4.1.2. Applying TR technique

Obtain the lower and upper approximations, relative roughness, mean roughness, total roughness of subsets of U based on attribute (LIP) with respect to attributes (MP, WP, CCI and S) are given in Herawan *et al.*, in 2010 [13] and the values of MMR and TR techniques can be summarized in Table 2.

Table 2. The Total Roughness of all Attributes in Table 1 using (TR) Technique

Attribute(with respect to)	Mean roughness				TR
MP	WP 0	LIP 0.25	CCI 0.1	S 0	0.0875
WP	MP 0	LIP. 0	CCI 0	S 0	0
LIP	MP 0.15	WP 0	CCI 0.1	S 0	0.0625
CCI	MP 0.15	WP 0	LIP 0.25	S 0.2	0.15
S	MP 0	WP 0	LIP 0	CCI 0.1	0.025

From Table 2, the value of TR technique of (LIP), *i.e.*, 0.0625 is lower than that of (MP), *i.e.*, 0.0875 and (CCI), *i.e.*, 0.15. Thus, the decision to select (LIP) as a clustering attribute is not appropriate, because the total roughness of attribute (LIP) is lower than that of attribute (CCI) in [13].

4.1.3. Applying (MMR) Technique

Getting the mean roughness of MMR by equation (6), next finding min-min of roughness MMR using equation (5) and the values of MMR technique summarized in Table 3 as in [13].

Table 3. The Minimum–minimum Roughness of all Attributes in Table 1 using MMR Technique

Attribute (with respect to)	**MMR Mean roughness**				**MMR**
MP	WP 1	LIP 0.75	CCI 0.9	S 1	0.75 0.9
WP	MP 1	LIP 1	CCI 1	S 1	1
LIP	MP 0.85	WP 1	CCI 0.9	S 1	0.85
CCI	MP 0.85	WP 1	LIP 0.75	S 0.8	0.75 0.8
S	MP 1	WP 1	LIP 1	CCI 0.9	0.9

The MMR of all attributes in Table 1 can be summarized as in Table 3. Herawan *et al.*, in 2010. [13] and from Table 3, the mean roughness of (MP) and (CCI) has the same minimum value, *i.e.*, 0.75. It has to look at the next lowest minimum value, and so on until the difference value is obtained. In this case, the CCI has the minimum value, *i.e.*, 0.8, as compared to (MP), *i.e.*, 0.9. Thus, the attribute (CCI) is selected as the clustering attribute.

4.1.4. Applying Maximum Dependency Attribute Algorithm

Calculating the degree of dependency of attribute (LIP) with respect to (MP, WP, CCI, & S) using equation(1) and applying the algorithm of MDA as in [13] and the values of maximum dependency of attributes summarized in Table 4.

Table 4. The Degree of Dependency of all Attributes in Table 1 using MDA Technique

Attribute (depends on)	Degree of dependency				MDA
MP	WP 0	LIP 0.5	CCI 0.2	S 0	0.5 0.2
WP	MP 0	LIP 0	CCI 0	S 0	0
LIP	MP 0.3	WP 0	CCI 0.2	S 0	0.3
CCI	MP 0.3	WP 0	LIP 0.5	S 0.4	0.5 0.4
S	MP 0	WP 0	LIP 0	CCI 0.2	0.2

From Table 4, the attributes (MP) and (CCI) has the same maximum degree of dependency, *i.e.*, 0.5. Based on the MDA algorithm, the next degree of attributes will be considered, until the tie is broken. In this case, the second degree corresponding to attribute CCI, *i.e.*, 0.4 is higher than that of MP, *i.e.*, 0.2. Therefore, attribute (CCI) is selected as the clustering attribute [13].

4.1.5. Applying our Maximum Significance Attribute Algorithm

We can get the significance of subsets of U based on attribute LIP with respect to attributes (MP, WP, CCI and S) via equation (7) and the results of the significance of all attributes can be summarized in Table 5.

a - The significance of attribute (LIP) with respect to attribute (MP), denoted as $\sigma_{MP}(LIP)$, can be calculated as follows.

Let C' all attributes except attribute MP
Where C' ={WP,LIP,CCI,S} And $C''=C'$ -{LIP}={WP,CCI,S}
U \ C' = {{1,3,9},{2,10},{4},{5},{6},{7},{8}},
U \ C'' = {{1,3,9},{2,10},{4},{5,6},{7},{8}}
U \ MP ={{1,2,4,5,7,9,10},{3,6,8}}

$$\sigma_{MP}(LIP) = \gamma_{C'}(MP) - \gamma_{C''}(MP) = \frac{7}{10} - \frac{5}{10} = 0.2$$

b –LIP with respect to WP
C' = {MP,LIP,CCI,S} , C'' = C' - {LIP} = {MP,CCI,S}
U \ C' = {{1,9},{2,5,10},{3,8},{2,7},{6}},
U \ C'' = {{1,9},{2,5,10},{3,8},{4,7},{6}}
U \ WP = {{1,3,5,6,7,9},{2,4,8,10}} ,

$$\sigma_{WP}(LIP) = \gamma_{C'}(WP) - \gamma_{C''}(WP) = \frac{3}{10} - \frac{3}{10} = 0$$

c – LIP with respect to CCI
C' = {MP,WP,LIP,S} , C'' = {MP,WP,S}

U \ C' = {{1,9},{2,5,10},{3,8},{4,7},{6}},
U \ C'' = {{1,7,9},{2,10},{3},{4},{5},{6},{8}}
U \ CCI = {{1,2,3,5,6,8,9,10},{4,7}}

$$\sigma_{CCI}(LIP) = \gamma_{C'}(CCI) - \gamma_{C''}(CCI) = \frac{10}{10} - \frac{7}{10} = 0.3$$

d – LIP with respect to S

C' = {MP,WP,LIP,CCI} , C'' = {MP,WP,CCI}
U \ C' = {{1,9},{2,10},{3,6},{4},{5},{7},{8}},
U \ C'' = {{1,5,9},{2,10},{3,6},{4},{7},{8}}
U \ S = {{1,3,4,7,8,9},{2,5,6,10}}

$$\sigma_{S}(LIP) = \gamma_{C'}(S) - \gamma_{C''}(S) = \frac{8}{10} - \frac{5}{10} = 0.3$$

Table 5. The Degree of Significance of all Attributes in Table 1 using MSA Technique

Attributes	Significance				MSA
MP	WP 0.2	LIP 0.2	CCI 0	S 0.2	0.2
WP	MP 0.1	LIP 0	CCI 0	S 0.1	0.1
LIP	MP 0.2	WP 0	CCI 0.3	S 0.3	0.3 0.3 0.2
CCI	MP 0	WP 0	LIP 0.3	S 0.3	0.3 0.3 0
S	MP 0.1	WP 0.1	LIP 0.3	CCI 0.2	0.3 0.2

From Table 5, the attribute (LIP) has the maximum significance of attributes, *i.e.*, 0.3 the next degree of attributes will be considered until the tie is broken. In this case, the second degree corresponding to attribute LIP, *i.e.*, 0.3 is same second degree of (MP), finally we get the third degree corresponding to attribute (LIP) *i.e.*, 0.2. In this case, the third degree corresponding to attribute (LIP), *i.e.*, 0.2 is higher than that of (CCI), *i.e.*, 0, Based on the MSA algorithm .Therefore, attribute (LIP) is selected as the clustering attribute.

5. The performance comparisons of MSA with that of BC, TR, MMR and MDA techniques

5.1. Objects splitting for TR, MMR and MDA techniques

For objects splitting, we use a divide-conquer method. For example, in Table 1 we can cluster (partition) the objects based on TR, MMR and MDA techniques which have the same clustering attribute (CCI) and similar objects splitting. Notice that, For first split we select first nearest attribute for the selecting clustering attribute induced by attribute LIP and the partition of the set of objects is{{1,3,6,8,9},{2,4,5,7,10}} and for second split we select the second nearest attribute from the select clustering attribute of TR, MMR and MDA algorithms is attribute (S). So we will redo split attribute (LIP) on attribute (S) with equivalence classes is {{1,3,4,7,8,9},{2,5,6,10}}. To this, we can split the objects using the hierarchical tree as follows.

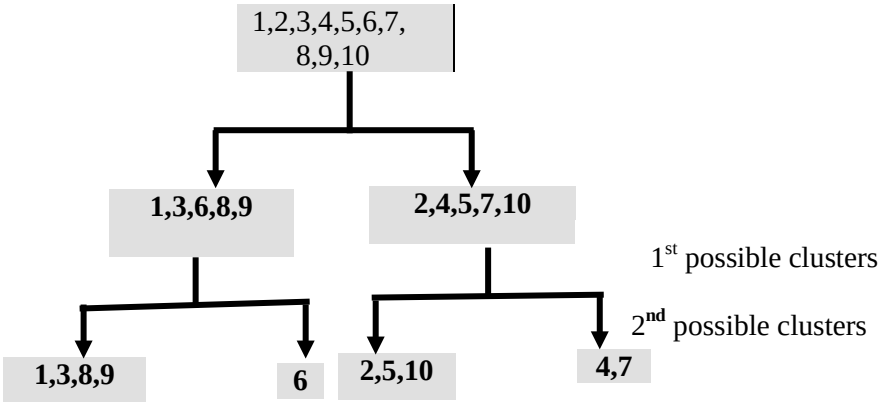

Figure 2. The Objects Splitting

5.2. The Purity Ratio for TR, MMR and MDA Techniques

The Acme company dataset contains 10 objects, where each data point represents information of a credit card in terms of 5 categorical attributes in the Acme company. The three techniques, where the total objects are divided into two classes so; we need to stop when we will get two clusters as only two credit cards, namely, (LIP) and (S) are described by five categorical attributes. The dataset comprises 5 objects for (LIP) and 4 for (S). Since there are two possible credit cards, the objects are split into two clusters. The results are summarized in Table 6. All of the 10 objects belong to the majority class label of the cluster in which they are classified. Thus, the overall purity of the cluster is 70%.

Table 6. The Overall Purity of MMR, TR and MDA

Clusters	C1	C2	Purity
Cluster 1	4	1	0.8
Cluster 2	3	2	0.6
Over all purity			0.7

5.3. Objects Splitting for MSA Technique

Splitting objects of the example, in Table 1 using the hierarchical tree based on the clustering attribute selected by using MSA is attribute (LIP). And for first split we select the first nearest attribute for the selecting clustering attribute induced by attribute (CCI) for the first split has the partition of the set of objects is{{1,2,3,5,6,8,9,10},{4,7}} and for second split we depend on attribute (S) where is the second nearest attribute from the selecting clustering attribute {{1,3,4,7,8,9},{2,5,6,10}}. So applying hierarchical tree of the objects as follows,

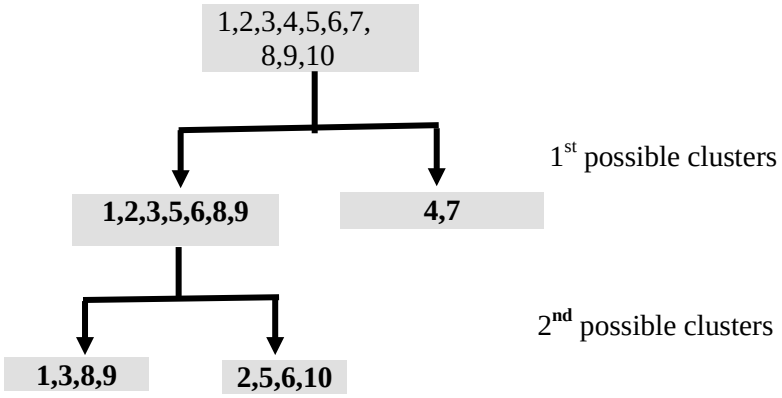

Figure 3. The MSA Objects Splitting

5.5. The Purity Ratio for MSA Technique

The Acme company dataset consists of 10 objects, where each data point represents information of a credit card in terms of five categorical attributes. Each credit card data point is classified into two classes. Therefore, for MSA, the split data is contained in two clusters. The results of applying the MSA algorithm to the Acme company dataset are summarized in Table 7, which gives the overall purity of the cluster as 75%.

Table 7. The Overall Purity of MSA Technique

Clusters	C1	C2	Purity
Cluster 1	4	4	0.5
Cluster 2	0	2	1
Over all purity			0.75

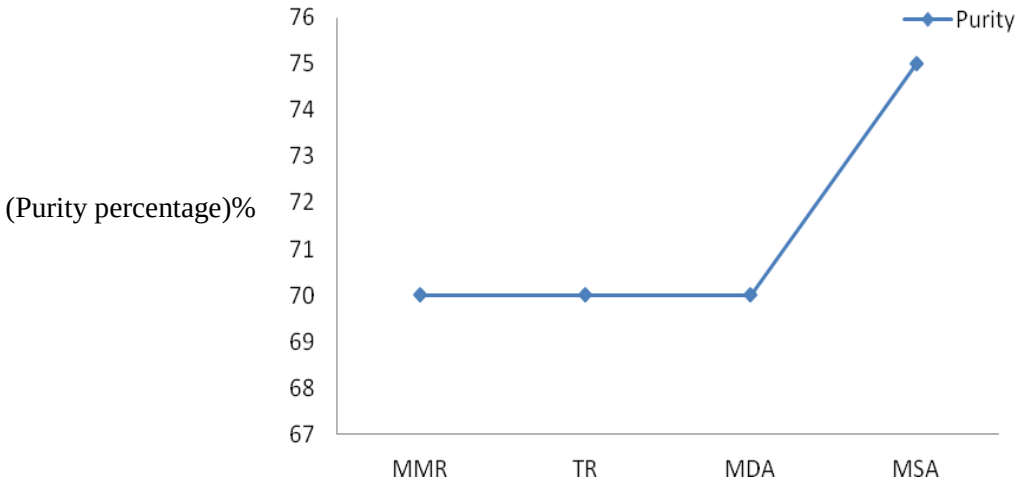

Figure 4. The Over All Purity

From Figure 4 we can see that the purity of selecting the clustering attribute using the (MMR, TR and MDA) algorithms is the same, *i.e.*, 70%, while that for the MSA algorithm is the highest of all the algorithms, *i.e.*, 75%.

6. Conclusions

In this paper, we proposed a new technique for selecting clustering attribute called (maximum significance of attributes) MSA. The proposed technique is based on rough set theory using the significance of attributes in information systems. The analysis of the MSA was presented in terms of purity ratio. The test case was selected, it showed that the MSA technique provides a convenient approach to higher clusters purity as compared to the four existing techniques. The proposed approach could also be applied in clustering data in large databases and *etc.*, we also experimented on some other conditional attribute tables with larger amount of data and drew the same conclusion. So the conclusion can be generalized.

Acknowledgement

This work was supported by the grant of mathematics department, faculty of science, Tanta University, Egypt.

References

[1] S. Wu, A. Liew, H. Yan and M. Yang, "Cluster analysis of gene expression data based on self-splitting and merging competitive learning", IEEE Transactions on Information Technology in BioMedicine, vol. 8, no. 1, **(2004)**, pp. 5-15.
[2] D. Jiang, C. Tang and A. Zhang, "Cluster analysis for gene expression data: A survey", IEEE Transactions on Knowledge and Data Engineering, vol. 16, no. 11, **(2004)**, pp. 1370-1386.
[3] K. Wong, D. Feng, S. Meikle and M. Fulham, "Segmentation of dynamic PET images using cluster Analysis", IEEE Transactions on Nuclear Science, vol. 49, no. 1, **(2002)**, pp. 200-207.
[4] S. Haimov, M. Michalev, A. Savchenko and O. Yordanov, "Classification of radar signatures by autoregressive model fitting and cluster analysis", IEEE Transactions on Geo Science and Remote Sensing, vol. 8, no. 1, **(1989)**, pp. 606-610.
[5] R. Mathieu and J. Gibson, "A methodology for large scale R&D planning based on cluster analysis", IEEE Transactions on Engineering Management, vol. 40, no. 3, **(1993)**, pp. 283-292.
[6] Z. Huang "Extensions to the k-means algorithm for clustering large data sets w categorical values", Data Mining and Knowledge Discovery, vol. 2, no. 3, **(1998)**, pp. 283-304.
[7] D. Gibson, J. Kleinberg and P. Raghavan, "Clustering categorical data: an approach based on dynamical systems", The Very Large Data Bases Journal, vol. 8, no. 3-4, **(2000)**, pp. 222-236.
[8] S. Guha, R. Rastogi and K. Shim, "ROCK: a robust clustering algorithm for categorical attributes", Information Systems, vol. 25, no. 5, **(2000)**, pp. 345-366.
[9] V. Ganti, J. Gehrke and R. Ramakrishnan, "CACTUS-clustering categorical data using summaries", Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, **(1999)**, pp. 73-83.
[10] A. Dempster, N. Laird and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm", Journal of the Royal Statistical Society, vol. 39, no. 1, **(1977)**, pp. 1-38.
[11] Z. Pawlak, "Rough Sets", International Journal of Computer and Information Science, vol. 11, no. 341, **(1982)**.
[12] Z. Pawlak and A. Skowron, "Rudiments of rough sets", Information Sciences, vol. 177, **(2007)**, pp. 3-27.
[13] T. Herawan, M. Deris and J. H. Abawajy, "A rough set approach for selecting clustering Attribute", Knowledge based systems, vol. 23, **(2010)**, pp. 220-231.
[14] L. J. Mazlack, A. He, Y. Zhu and S. Coppock, "A rough set approach in choosing clustering attributes", Proceedings of the ISCA 13th, International Conference (CAINE-2000), **(2000)**, pp. 1-6.
[15] Z. Pawlak, "Rough Sets: Theorielial Aspects of Reasoning about data, System theory", knowledge Engineering and Problem Solver Klumer Academic Publishers, Dordrecht, vol. 9, **(1991)**.
[16] R. J. Roiger and M. W. Geatz, "Data Mining: A Tutorial-Based Primer", Addison Wesley, **(2003)**.
[17] T. Herawan, R. Ghazali, I. Yanto and M. Deris, "Rough Set Approach for Categorical Data Clustering", International Journal of Database Theory and Application, vol. 3, no. 1, **(2010)**, pp. 33-52.
[18] Z. Pawlak, "Rough Sets: Theorielial Aspects of Reasoning about data System theory", knowledge Engineering and Problem Solver Klumer Academic Publishers, Dordrecht, vol. 9, **(1991)**.

[19] D. Parmar, T. Wu and J. Blackhurst, "MMR: an algorithm for clustering categorical data using rough set theory", Data and Knowledge Engineering, vol. 63, **(2007)**, pp. 879-893.

Authors

W. A. Hassanein, is a lecturer of Mathematical Statistics - Mathematics Department and a Faculty of Science- Tanta University- Egypt.

Amr A. Elmelegy, he is a M.Sc. candidate in Data Cluster Analysis at scholarship in faculty of science, Tanta University, Egypt. His research area includes clustering categorical data on rough set theory.

