European Scientific Journal January 2014 edition vol.10, No.3 ISSN: 1857 – 7881 (Print) e - ISSN 1857- 7431

CLUSTERING ALGORITHMS FOR CATEGORICAL DATA USING CONCEPTS OF SIGNIFICANCE AND DEPENDENCE OF ATTRIBUTES

Wafa Anwar Hassanein, PhD
Amr Ahmed Elmelegy, MA
Mathematics Department, Faculty of Science,
Tanta University, Tanta, Egypt

Abstract

Clustering categorical data is an essential and integral part of data mining. In this paper, we propose two new algorithms for clustering categorical data, namely, the Standard Deviation of Standard Deviation Significance and Standard Deviation of Standard Deviation Dependence algorithms. The proposed techniques are based mainly on rough set theory, taking into account the significance and dependence of attributes of database concepts. Analysis of the performance of the proposed algorithms compared with others shows their efficiency as well as ability to handle uncertainty together with categorical data.

Keywords: Clustering, Rough set theory, Dependence of attributes, Significance of attributes, Standard deviation

1. Introduction

Imagine that you are given a set of data objects for analysis where, unlike in classification, the class label of each object is unknown. This is quite common in large databases, because assigning class labels to a large number of objects can be a very costly process. Clustering is the process of grouping the data into classes or clusters, so that objects within a cluster have high similarity to one another, but are very dissimilar to objects in other clusters (Han J 2006). Dissimilarities are assessed based on the attribute values describing the objects; distance measures are frequently used for this purpose. Clustering has its roots in many areas, including data mining, statistics, biology, and machine learning.

Clustering techniques are used in many areas including manufacturing, medicine, nuclear science, radar scanning, and research and

development planning. For example (Wu S, Liew A, Yan H, Yang M 2004) developed a clustering algorithm specifically designed for handling the complexity of gene data, while (Jiang D, Tang C, Zhang A 2004) analyzed a variety of cluster techniques that can be applied to gene expression data. (Wong K, Feng D, Meikle S, Fulham M 2002) presented an approach for segmenting tissues in a nuclear medical imaging method known as positron emission tomography. (Haimov S, Michalev M, Savchenko A, Yordanov O 1989) used cluster analysis to segment radar signals in scanning land and marine objects. Finally, (Mathieu R, Gibson J 1993) used cluster analysis as part of a decision support tool for large scale research and development planning to identify programs in which to participate and to determine resource allocation.

The problem with all the algorithms mentioned above is that they mostly deal with numerical datasets, that is, databases with attributes with numeric domains. The main advantage of dealing with numerical attributes is that they are very easy to handle and moreover, it is easy to define similarity between them. On the other hand, categorical data have multi-valued attributes. Thus, similarity can be defined as common objects, common values of attributes, or an association between the two. In such cases, horizontal co-occurrences (common value of objects) as well as vertical co-occurrences (common value of attributes) must be examined (Wu S, Liew A, Yan H, Yang M 2004).

Various algorithms that can handle categorical data have been proposed, including the work by (Parmar D, Wu T, Blackhurst J 2007, Gibson D, Kleinberg J, Raghavan P 2000, Jiang D, Tang C, Zhang A 2004 and Dempster A P, Laird N M, Rubin D B 1977). While these algorithms and methods are very useful in forming clusters from categorical data, they have the disadvantage of being unable to deal with uncertainty. However, in real-world applications it has been found that there is often no sharp boundary between clusters. Recently, work has been done by Huang (Huang Z 1998) and Kim et al. (Kim D, Lee K, Lee D 2004) in developing several clustering algorithms using fuzzy sets, which can handle categorical data. However, these algorithms suffer from stability problems as they do not provide satisfactory values owing to multiple executions of the algorithms.

There is, therefore, a need for a robust algorithm that can handle uncertainty together with categorical data. For categorical data, fewer algorithms are available for grouping objects with similar characteristics. Furthermore, some of these have a complicated clustering process, while others have stability issues. Nevertheless, the results of all of these algorithms have low purity. In 2007, the Minimum-Minimum Roughness algorithm was proposed (Parmar D, Wu T, Blackhurst J 2007), which uses rough set theory concepts to deal with the above problems in clustering

categorical data. Later, in 2009, this algorithm was further improved to create the Minimum Mean Roughness (MMeR) algorithm (Tripathy B K , Prakash M S 2009), which can handle hybrid data. More recently in 2011, MMeR was again improved to develop an algorithm called the Standard Deviation Roughness (SDR) (Tripathy B K, Ghosh A 2011), which can also handle hybrid data. Later in 2011, SDR was further improved to create a new algorithm, the Standard Deviation of Standard Deviation Roughness (SSDR) (Tripathy B K, Ghosh 2011).

In this paper we propose two new algorithms that can deal with both uncertainty and categorical data at the same time, and which are better than their predecessors, SDR and SSDR. These algorithms are called the Standard Deviation of Standard Deviation Significance (SSDS) and the Standard Deviation of Standard Deviation Dependence (SSDD). The proposed techniques are based on rough set theory and take into account the significance (S) and dependence (D) of attributes in the database. An analysis and comparison of the performance of the SDR, SSDR, SSDS, and SSDD algorithms shows that SSDD has the highest purity ratio of the algorithms in this and previous series. To establish the superiority of this algorithm we tested all algorithms on the ACME company dataset.

The rest of this paper is organized as follows. **Section 2** presents some important definitions. In **Section 3** we discuss the SSDS and SSDD algorithms. **Section 4** describes the experimental setup, while **Section 5** compares the performance of SSDS and SSDD with previous algorithms. **Section 6** presents the conclusions of this work.

2. Main Concepts

Definition 1 (Information system)

In a rough set, information systems are used to represent knowledge. An information system is denoted as $S=(U,A,V,F)$, where U is a non empty finite set of objects, A is a non empty finite set of attributes, $V=\bigcup_{a\in A}V_a$, where V_a is the domain (value set) of attribute a, and $f:U\times A\to V$ is a total function, called the information (knowledge) function, such that $f(u,a)\in V_a$ for every $(u,a)\in U\times A$ (Pawlak Z, Skowron A 2007).

Definition 2 (Indiscernibility relation)

Indiscernibility relation (*Ind* (*B*)): *Ind* (*B*) is a relation on *U*. 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 $a\in B$ as proposed in (Pawlak Z,

Skowron A 2007). That is, $x_i, x_j \in Ind(B)$ if and only if $\forall a \in B$, where $B \subseteq A$ and $a(x_i) = a(x_j)$.

Definition 3 (Equivalence classes)

Equivalence class ($[x_i]_{Ind(B)}$): Given *Ind* (*B*), the set of objects x_i with the same values for the set of attributes in *B* comprises the equivalence class, $[x_i]_{Ind(B)}$, also known as an elementary set with respect to B.

Definition 4 (Upper approximation)

Given the set of attributes *B* in *A* and the set of objects *X* in *U*, the upper approximation of *X* is defined as the union of all the elementary sets contained in *X*. That is,

$$\overline{X}_B = \bigcup \{X_i \mid [X_i]_{Ind(B)} \cap X \neq \phi\}. \quad (1)$$

Definition 5 (Lower approximation)

Given the set of attributes *B* in *A* and set of objects *X* in *U*, the lower approximation of *X* is defined as the union of all the elementary sets contained in *X*. That is,

$$\underline{X}_B = \bigcup \{X_i \mid [X_i]_{Ind(B)} \subseteq X\}. \quad (2)$$

Definition 6 (Roughness)

The ratio of the cardinality of the lower approximation and that of the upper approximation is defined as the accuracy of estimation, which is a measure of roughness. In (Tripathy B K, Ghosh 2011) this is defined as

$$R_B(X) = 1 - \frac{|\underline{X}_B|}{|\overline{X}_B|}. \quad (3)$$

If $R_B(X)$, *X* is crisp with respect to *B*; in other words, *X* is precise with respect to *B*. If $R_B(X) < 1$, *X* is rough with respect to *B*; that is, *B* is vague with respect to *X*.

Definition 7 (Relative roughness)

Given that $a_i \in A$, *X* is a subset of objects with one specific value α of attribute a_i, and $\underline{X_{a_j}(a_j = \alpha)}$ and $\overline{X_{a_j}(a_j = \alpha)}$ denote, respectively, the lower and upper approximations of *X* with respect to $\{a_j\}$, then $R_{a_j}(X)$ is

defined as the roughness of *X* with respect to $\{a_j\}$ as given in (Tripathy B K, Ghosh 2011):

$$R_{a_j}(X / a_j = \alpha) = 1 - \frac{\left|\underline{X}_{a_j}(a_j = \alpha)\right|}{\left|\overline{X}_{a_j}(a_j = \alpha)\right|}, \text{ where } a_i, a_j \in A, \text{ and } a_i \neq a_j . \quad (4)$$

Definition 8 (Mean roughness) (Tripathy B K, Ghosh 2011)

Let *A* have *n* attributes with $a_i \in A$ and *X* be the subset of objects with a specific value α of attribute a_i. Then we define the mean roughness for the equivalence class $a_i = \alpha$, denoted as $MeR(a_i = \alpha)$ (Tripathy B K, Ghosh A 2011), As

$$MeR(a_i = \alpha) = (\sum_{\substack{j=1 \\ j \neq i}}^{n} R_{a_j}(X / a_j = \alpha) / (n-1)) . \quad (5)$$

Definition 9 (Standard deviation of roughness)

After calculating the mean of each $a_i \in A$, we apply the standard deviation to each a_i using the formula defined in (Tripathy B K, Ghosh 2011):

$$SD(a_i = \alpha) = \sqrt{(1/(n-1))\sum_{i=1}^{n-1}(R_{a_j}(X / a_j = \alpha) - MeR(a_i = \alpha))^2} \quad (6)$$

Definition 10 (Significance of attribute)

Let the significance of attribute $a_i \in$ A related to $a_j \in A$, where *A* represents all the attributes, denoted by $S_{a_j}(a_i)$, be $S_{a_j}(a_i)$ = $\sigma_{a_j}(a_i) = \gamma_{A'}(a_j) - \gamma_{A''}(a_j)$, (7)

where $A' = A - \{a_j\}$, $A'' = A' - \{a_i\}$ (see Pawlak Z 1991).

Definition 11 (Mean significance)

The mean significance of attribute $a_i \in A$ with respect to attribute $a_j \in A$, where $i \neq j$, denoted as $Significance_{a_j}(a_i)$, is evaluated as

$$MeS(a_i) = \frac{\sum_{\substack{j=1 \\ j \neq i}}^{n} S_{a_j}(a_i)}{n-1} \quad . \tag{8}$$

Definition 12 (Dependence of an attribute)

Suppose $S = (U,A,V,F)$ is an information system and a_i and a_j are subsets of A. The dependence of attribute a_i on a_j with degree k ($0 < k < 1$), is denoted as $a_i \Rightarrow a_j$. Degree k proposed by Herawan et al. (Herawan T, Deris M, Abawajy J H 2010), is defined as

$$K = \gamma_{a_j}(a_i) = \frac{\sum_{X \in U/a_i} \left| a_j(X) \right|}{|U|} \quad . \tag{9}$$

Definition 13 (Mean dependence)

The mean dependence of attribute $a_i \in A$ with respect to attribute $a_j \in A$, where $i \neq j$, denoted by $MeD(a_i)$, is calculated as:

$$MeD(a_i) = \frac{\sum_{\substack{j=1 \\ j \neq i}}^{n} \gamma_{a_j}(a_i)}{n-1} \, . \tag{10}$$

Definition 14 (Standard deviation of dependence)

After calculating the mean dependence of each $a_i \in$A, we apply the standard deviation to each a_i using the formula:

$$SD(a_i) = \sqrt{(1/(n-1)) \sum_{j=1}^{n-1} (\gamma_{a_j}(a_i) - MeD(a_i))^2} \quad . \tag{11}$$

Definition 15 (Standard deviation of significance)

After calculating the mean significance of each $a_i \in$A, we apply the standard deviation to each a_j using the formula:

$$SD(a_i) = \sqrt{(1/(n-1))\sum_{j=1}^{n-1}(S_{a_j}(a_i) - MeS(a_i))^2} \quad . \qquad (12)$$

Definition 16 (Purity ratio)

To compare the SDR, SSDR, SSDS, and SSDD algorithms, which can all handle uncertainty and categorical data together, we developed a measure of purity. The traditional approach for calculating the purity of a cluster proposed in (Herawan T, Ghazali R, Yanto I, Deris M 2010) is given as

$$purity = \frac{number\ of\ occurences\ \text{in}\ both\ the\ cluster\ and\ its\ corresponding\ class}{number\ of\ data\ sets} \qquad (13)$$

$$\text{over all purity} = \frac{\sum_{i=1}^{\#of\ clusters} purity(i)}{\#of\ clusters} \qquad (14)$$

3. Proposed Algorithms

Having introduced the notations and definitions of the concepts in the previous section, here we present our algorithms.

Algorithm 1 for SSDS

1. Procedure SSDS(U,k)
2. Begin
3. Set current number of clusters, CNC = 1
4. Set ParentNode = U
5. Loop1:
6. If CNC < k and CNC ≠ 1 then
7. ParentNode = Proc ParentNode(CNC)
8. End if
   // Clustering the ParentNode
9. For each $a_i \in A$ (i = 1 to *n*, where *n* is the number of attributes in *A*)
10. Determine $[X_m]_{Ind(a_i)}$ (*m*=1 to number the objects)
11. Get $U/ind(A')$, where A' denotes the family (*A*) except $\{a_j\}$ of all equivalence classes of $A - a_j$, written as U/A'.
12. Get U/A'', which denotes the equivalence classes of $\{A-\{a_j\}\}-\{a_i\}$ or A' - $\{a_i\}$.
13. Get U/a_j.
14. Get $pos_{A'}(a_j)$.

15. Compute $\gamma_{A'}(a_j)$, which is the dependence of a_j on condition attribute set A'.
16. Get $pos_{A''}(a_j)$.
17. Compute $\gamma_{A''}(a_j)$.
28. Compute $\sigma_{a_j}(a_i)$.

18. Next
19. $$MeS(a_i) = \frac{\sum_{\substack{j=1 \\ j \neq i}}^{n} S_{a_j}(a_i)}{n-1}$$
20. Next
21. Apply standard deviation significance
$$SDS(a_i) = \sqrt{(1/(n-1))\sum_{j=1}^{n-1}(S_{a_j}(a_i) - MeS(a_i))^2}$$
22. Next
23. Mean of all $SDS(a_i)$
$$MeSDS(a_i) = \frac{\sum_{i=1}^{n} MinSDS(a_i)}{n}$$
24. Next
25. Apply standard deviation of standard deviation significance
$$SSDS = \sqrt{(1/(n-1))\sum_{i=1}^{n}(SDS(a_i) - MeSDS(a_i))^2}$$
26. Determine splitting attribute a_i corresponding to the standard deviation significance
27. Perform binary split on the splitting attribute a_i

Algorithm 2 for SSDD

1. Procedure SSDD(U,k)
2. Begin
3. Set current number of clusters, CNC = 1
4. Set ParentNode = U
5. Loop1:
6. If CNC < k and CNC ≠ 1 then
7. ParentNode = Proc ParentNode(CNC)
8. End if
   // Clustering the ParentNode
9. For each $a_i, a_j \in A$ (i, j = 1 to n, where n is the number of attributes in A)
10. Determine U / a_i , U / a_j
11. Calculate $K = \gamma_{a_j}(a_i) = \dfrac{\sum_{X \in U/a_i} \left| a_j(X) \right|}{|U|}$
12. Next
13. $MeD(a_i) = \dfrac{\sum_{\substack{j=1 \\ j \neq i}}^{n} \gamma_{a_j}(a_i)}{n-1}$
14. Next
15. Apply standard deviation dependence
$$SDD(a_i) = \sqrt{(1/(n-1))\sum_{j=1}^{n-1}(\gamma_{a_j}(a_i) - MeD(a_i))^2}$$
16. Next
17. Mean of all SDD(a_i)
$$\text{MeSDD}(a_i) = \frac{\sum_{i=1}^{n} SDD(a_i)}{n}$$
18. Next
19. Apply standard deviation of standard deviation dependence
$$SSDD = \sqrt{(1/(n-1))\sum_{i=1}^{n-1}(SDD(a_i) - MeSDD(a_i))^2}$$
20. Determine splitting attribute a_i corresponding to the standard deviation dependence
21. Perform binary split on the splitting attribute a_i

4. Experiments

For the experiment, we used the credit card promotion dataset given in (Roiger R J , Geatz M W 2003), a portion of which is listed in Table 1. There are five categorical attributes (n = 5): magazine promotion (MP), watch promotion (WP), life insurance promotion (LIP), credit card insurance (CCI), and sex (S). Each of the attributes has two distinct values, (l = 2), i.e., yes or no, and ten objects (m = 10) are considered.

For the computations, we consider the information system shown in Table 1.

Table 1 A subset of the credit card promotion dataset from Acme Credit Card Company database (Roiger R J , Geatz M W 2003)

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

4.1.1. Obtain 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. Apply SDR and SSDR algorithms

Calculations to obtain the lower and upper approximations, relative roughness, mean roughness, and standard deviation of the roughness of

subsets of U based on attribute LIP with respect to attributes MP, WP, CCI, and S are given below.

4.1.2.1.Obtain lower and upper approximations

a) LIP with respect to MP

$\underline{X(LIP=yes)}=\phi$, $\overline{X(LIP=yes)}=\{1,2,4,5,7,9,10\}$

$\underline{X(LIP=no)}=\{3,6,8\}$, $\overline{X(LIP=no)}=\{1,2,3,4,5,6,7,8,9,10\}$

b) LIP with respect to WP

$\underline{X(LIP=yes)}=\phi$, $\overline{X(LIP=yes)}=\{1,2,3,4,5,6,7,8,9,10\}$

$\underline{X(LIP=no)}=\phi$, $\overline{X(LIP=no)}=\{1,2,3,4,5,6,7,8,9,10\}$

c) LIP with respect to CCI

$\underline{X(LIP=yes)}=\{4,7\}$, $\overline{X(LIP=yes)}=\{1,2,3,4,5,6,7,8,9,10\}$

$\underline{X(LIP=no)}=\phi$, $\overline{X(LIP=no)}=\{1,2,3,5,6,8,9,10\}$

d) LIP with respect to S

$\underline{X(LIP=yes)}=\phi$, $\overline{X(LIP=yes)}=\{1,2,3,4,5,6,7,8,9,10\}$

$\underline{X(LIP=no)}=\phi$, $\overline{X(LIP=no)}=\{1,2,3,4,5,6,7,8,9,10\}$

4.1.2.2. Relative roughness

a) LIP with respect to MP roughness

$R_{MP}(X \mid LIP=yes)=1-0 = 1,$ $R_{MP}(X \mid LIP=no)=1-0.3=0.7$

b) LIP with respect to WP roughness

$R_{WP}(X \mid LIP=yes)=1-0 = 1,$ $R_{WP}(X \mid LIP=no)=1-0 = 1$

c) LIP with respect to CCI roughness

$R_{CCI}(X \mid LIP=yes)=1-0.2 = 0.8,$ $R_{CCI}(X \mid LIP=no)=1-0 = 1$

d) LIP with respect to S roughness

$R_{S}(X \mid LIP=yes)=1-0 = 1,$ $R_{S}(X \mid LIP=no)=1-0 = 1$

4.1.2.3. Mean roughness (MeR)

a) MeR(LIP=yes) = $\frac{1+1+0.8+1}{4} = 0.95$

b) MeR(LIP=no) = $\frac{0.7+1+1+1}{4} = 0.925$

4.1.2.4. Standard deviation of roughness

a) SD(LIP=yes) $= \sqrt{\frac{1}{4}\times((1-0.95)^2+(1-0.95)^2+(1-0.95)^2+(0.8-0.95)^2)^2}$

$= 0.0866$

b) SD(LIP=no) $= \sqrt{\frac{1}{4}\times((1-0.925)^2+(1-0.925)^2+(1-0.925)^2+(0.7-0.925)^2)^2}$

$= 0.12$

A similar process is followed, changing the value of α (for α = 'yes' or 'no') and keeping the value of a_i constant. Finally, we obtain two standard deviation values for each α, which are stored in variables. After calculating the standard deviation (SD) of roughness for each α, we take the minimum of these values for α and store this in another variable.

The above procedure is carried out for each a_i (that is, a_i = 'MP', 'WP', 'CCI', and 'S'), and the corresponding values are stored in variables. After completing the above step we use the minimum values for the next calculation. We apply the SSDR standard deviation to the minimum values to obtain the splitting attributes. If the value of SSDR does not match any of the minimum values, we take the closest minimum value as the splitting attribute and perform binary splitting. In other words, we divide the table into two clusters; the results of applying SDR and SSDR to all attributes listed in Table 1 are summarized in Table 2.

4.1.3. Apply the SSDS algorithm

We obtain the significance, mean significance, and standard deviation of the significance of attribute LIP with respect to attributes MP, WP, CCI, and S as discussed below.

4.1.3.1.Obtain the concept of significance of an attribute

a) The degree of significance of attribute LIP with respect to attribute MP, denoted as $\sigma_{MP}(LIP)$, can be calculated as follows. Let C' denote all attributes except attribute MP; thus

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'' = {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}}

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

4.1.3.2. Find the mean significance

$$MeS(LIP) = \frac{0.2+0+0.3+0.3}{4} = 0.2$$

4.1.3.3 Obtain standard deviation of significance SDS(a_i)

$$\text{SDS(LIP)} = \sqrt{\frac{1}{4} \times ((0.2-0.2)^2 + (0-0.2)^2 + (0.3-0.2)^2 + (0.3-0.2)^2)^2}$$
$$= 0.12$$

These steps are typically carried out on the other attributes, keeping the value of a_i constant for (a_i = 'MP', 'WP', 'CCI', and 'S'). After calculating the standard deviation of significance (SDS) of each a_i, we finally obtain one standard deviation value for each attribute a_i; these values are stored in a variable.

In the next calculation, we apply the SSDS to these values to obtain the splitting attributes. If the value of SSDS does not match any of the values of SDS exactly, we take the closest minimum value as the splitting attribute

and perform binary splitting; that is, we divide this table into two clusters. The SDS and SSDS results for all attributes are summarized in Table 3.

4.1.4.Apply the SSDD algorithm

We obtain the concept of dependence, the mean dependence, and the standard deviation of the dependence of attribute LIP with respect to attributes MP, WP, CCI, and S as given below.

4.1.4.1.Obtain the concept of dependence of an attribute

From Table 1, for each attribute, there are five partitions of U induced by indiscernibility relations on each attribute. The degree of dependence of attribute LIP on attribute MP, denoted $MP \Rightarrow LIP$, is calculated as follows.

a) $MP \Rightarrow LIP$

$$K = \frac{\sum_{X \in U/LIP} \left| \underline{MP(X)} \right|}{|U|} = \frac{|\{3,6,8\}|}{|\{1,2,3,4,5,6,7,8,9,10\}|} = \frac{3}{10} = 0.3$$

In the same way, we obtain the following:

b) $WP \Rightarrow LIP$

$$K = \frac{\sum_{X \in U/LIP} \left| \underline{WP(X)} \right|}{|U|} = \frac{|\{\ \}|}{|\{1,2,3,4,5,6,7,8,9,10\}|} = \frac{0}{10} = 0$$

c) $CCI \Rightarrow LIP$

$$K = \frac{\sum_{X \in U/LIP} \left| \underline{CCI(X)} \right|}{|U|} = \frac{|\{4,7\}|}{|\{1,2,3,4,5,6,7,8,9,10\}|} = \frac{2}{10} = 0.2$$

d) $S \Rightarrow LIP$

$$K = \frac{\sum_{X \in U/LIP} \left| \underline{S(X)} \right|}{|U|} = \frac{|\{\ \}|}{|\{1,2,3,4,5,6,7,8,9,10\}|} = \frac{0}{10} = 0$$

4.1.4.2.Find the mean dependence

$$MeD(LIP) = \frac{0.3+0+0.2+0}{4} = 0.125$$

4.1.4.3.Calculate the standard deviation of the dependence SDD(a_i)

$$SDD(LIP) = \sqrt{\frac{1}{4} \times ((0.3-0.125)^2 + (0-0.125)^2 + (0.2-0.125)^2 + (0-0.125)^2)} = 0.19$$

The above procedure is repeated changing the attribute and keeping the value of a_i constant (for a_i ='MP', 'WP', 'CCI', and 'S'). Finally we obtain one standard deviation value for each attribute a_i, and again these values are stored in variables. The corresponding values are also stored in variables. After completing this step, the stored values are used in the next calculation in which we apply the second standard deviation of the standard deviation dependence to the values to obtain the splitting attributes. If the value of the second algorithm (SSDD) does not match any of the values of SDD we take the closest minimum value as the splitting attribute and perform binary splitting; that is, we again divide the table into two clusters. The results of SDD and SSDD are summarized in Table 4.

4.2. Results of the algorithms

In this section we present the original results tested on the ACME company dataset, obtained using the SDR and SSDR algorithms (as shown in Table 2). We also give our results for the SSDS and SSDD algorithms, given in Tables 3 and 4, respectively.

Table 2 Experimental results for SDR and SSDR algorithms

Attributes	MeR($a_i = \alpha$)		SD($a_i = \alpha$)		SDR	SSDR
	MeR($a_i = yes$)	MeR($a_i = no$)	SD($a_i = yes$)	SD($a_i = no$)		
MP	0.825	1	0.2	0	0 0.2	0.035
WP	1	1	0	0	0	
LIP	0.95	0.925	0.08	0.12	0.08	
CCI	1	0.6	0	0.11	0 0.11	
S	0.95	1	0.086	0	0 0.086	

Table 3 Experimental results for SSDS algorithm

Attributes	Significance				$MeS(a_i)$	SDS	SSDS
MP	WP 0.2	LIP 0.2	CCI 0	S 0.2	0.15	0.0866	0.038
WP	MP 0.1	LIP 0	CCI 0	S 0.1	0.05	0.05	
LIP	MP 0.2	WP 0	CCI 0.3	S 0.3	0.2	0.12	
CCI	MP 0	WP 0	LIP 0.3	S 0.3	0.15	0.15	
S	MP 0.1	WP 0.1	LIP 0.3	CCI 0.2	0.175	0.08	

Table 4 Experimental results for SSDD algorithm

Attribute (depends on)	Degree of dependence				$MeD(a_i)$	$SDD(a_i)$	$SSDD(a_i)$
MP	WP 0	LIP 0.5	CCI 0.2	S 0	0.175	0.2	0.087
WP	MP 0	LIP 0	CCI 0	S 0	0	0	
LIP	MP 0.3	WP 0	CCI 0.2	S 0	0.125	0.19	
CCI	MP 0.3	WP 0	LIP 0.5	S 0.4	0.3	0.187	
S	MP 0	WP 0	LIP 0	CCI 0.2	0.05	0.0866	

5. Performance comparison of the SDR, SSDR, SSDS, and SSDD algorithms

5.1. Object splitting in SDR, SSDR, and SSDS algorithms

For object splitting, we use a divide-conquer method. For the example in Table 1, we can cluster (partition) the objects using the SDR, SSDR, and SSDS algorithms, which have the same clustering attribute and similar object splitting, i.e., WP. Note that the partition of the set of objects for the first split induced by attribute WP is {{1,3,5,6,7,9},{2,4,8,10}}, while for the second split, we select the second closest attribute from the selected clustering attribute of the SDR, SSDR, and SSDS algorithms, which is attribute S. Thus, we redo the split attribute WP on attribute S with equivalence classes {{1,3,4,7,8,9},{2,5,6,10}}. We therefore, split the objects according to the hierarchical tree shown in Fig. 1.

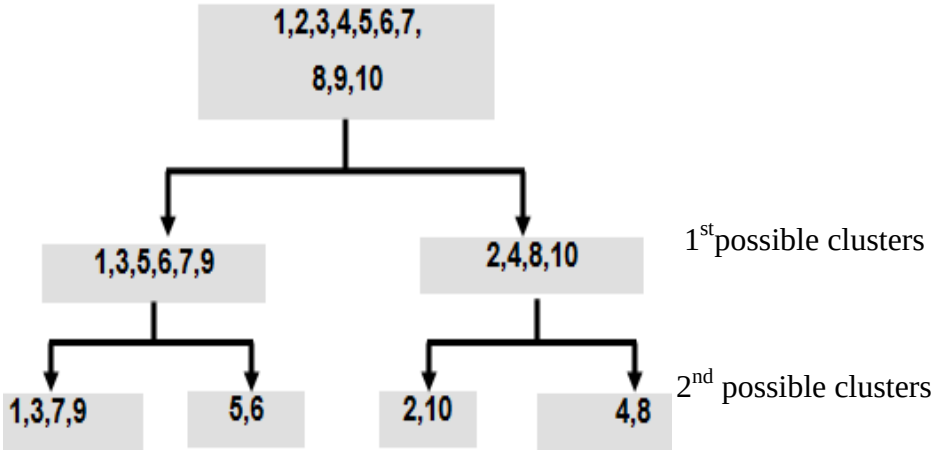

Fig. 1 Object splitting in SDR, SSDR, and SSDS

5.2.Purity ratios of SDR, SSDR, and SSDS algorithms

The Acme company dataset contains ten objects, where each data point represents information of a credit card in terms of five categorical attributes in the Acme company. The three algorithms SDR, SSDR, and SSDS have the same classification for each, with the objects divided into two classes, Thus, we need to stop when we get two clusters as only two credit cards, namely, watch promotion and sex, are described by the five categorical attributes. The dataset comprises six objects for watch promotion (WP) and sex (S). Since there are two possible credit cards, the objects are split into two clusters. The results are summarized in Table 5. All the ten objects belong to the majority class label of the cluster in which they are classified. Thus, the overall purity of the cluster is 58.33%.

Table 5 Overall purity of SDR, SSDR, and SSDS algorithms

Clusters	C 1	C2	Purity
Cluster 1	4	2	4/6
Cluster 2	2	2	2/4
Overall purity			0.5833

5.3.Object splitting in SSDD algorithm

Attribute S is used for splitting objects in the first split of the example given in Table 1 using the hierarchical tree based on the clustering attribute selected by SSDD. This split partitions the set of objects into {{1,3,4,7,8,9},{2,5,6,10}}. For the second split we depend on attribute CCI, which is the second closest attribute to the selected clustering attribute {{1,2,3,5,6,8,9,10},{4,7}}. The hierarchical tree for splitting the objects is shown in Fig. 2.

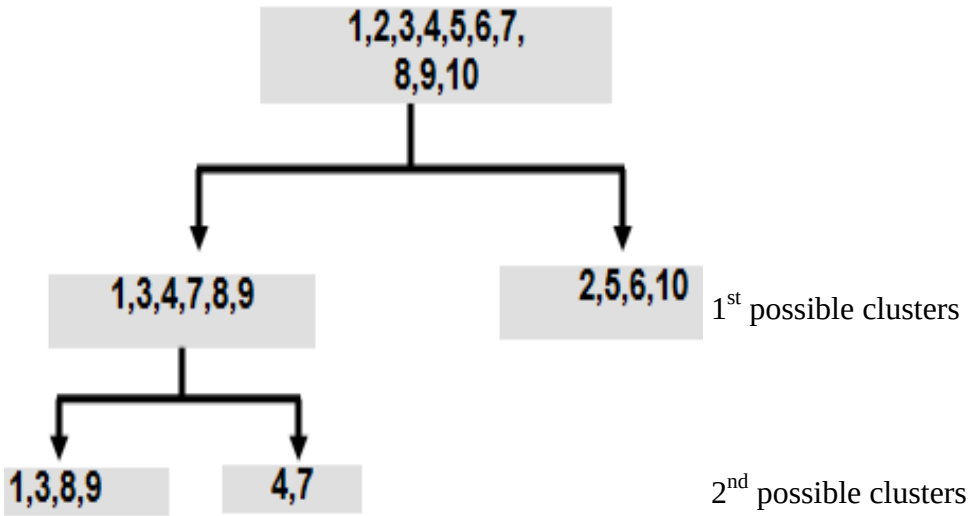

Fig. 2 Object splitting in SSDD algorithm

5.4. Purity ratio of SSDD algorithm

The Acme company dataset consists of ten 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 SSDD, the split data is contained in two clusters. The results of applying the SSDD algorithm to the Acme company dataset are summarized in Table 6, which gives the overall purity of the cluster as 83.33%.

Table 6 Overall purity of SSDD algorithm

Clusters	C 1	C2	Purity
Cluster 1	4	2	4/6
Cluster 2	0	4	1
Overall purity			0.8333

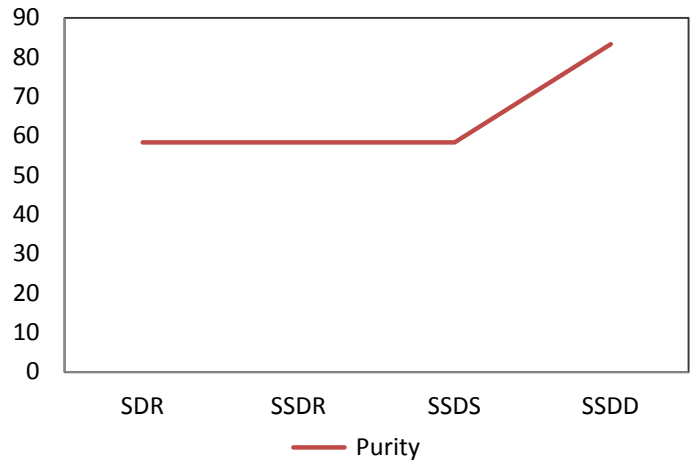

Fig. 3 Comparison of overall purity

From Fig. 3 we can see that the purity of selecting the clustering attribute using the SDR, SSDR, and SSDS algorithms is the same, i.e., 58.33%, while that for the SSDD algorithm is the highest of all the algorithms, i.e., 83.33%.

6. Conclusion

In this paper, we proposed two new algorithms for obtaining the splitting clustering attributes, that is, the SSDS and SSDD algorithms. The proposed techniques are based on rough set theory using the significance of attributes in information systems and the dependence of attributes in the database. Analysis of a test case shows that using the SSDS algorithm provides an easier method compared with the SDR and SSDR algorithms while yielding the same purity. Using the SSDD algorithm yields the highest

purity compared with the other algorithms. The proposed approach can also be used for clustering data in large databases. We also carried out an experiment with various other conditional attribute tables with larger amounts of data and obtained similar results. Thus, our conclusion can be generalized.

Acknowledgement

This work was supported by a grant from the Mathematics Department, Faculty of Science, Tanta University, Egypt.

References:

Dempster A P., Laird N M., Rubin D B. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society 39(1) (pp. 1–38) 1977.

Gibson D., Kleinberg J., Raghavan P. Clustering categorical data: An approach based on dynamical systems. The Very Large Data Bases Journal 8(3–4) (pp. 222–236) 2000.

Jiang D., Tang C., Zhang A. Cluster analysis for gene expression data: A survey. IEEE Transactions on Knowledge and Data Engineering 16(11) (pp. 1370–1386) 2004.

Han J. Kamber M Data mining: Concepts and techniques, 2nd Ed. Morgan Kaufmann Publishers, Burlington MA 2006.

Haimov S., Michalev M., Savchenko A., Yordanov O. Classification of radar signatures by autoregressive model fitting and cluster analysis. IEEE Transactions on Geo Science and Remote Sensing 8(1) (pp. 606–610) 1989.

Herawan T., Deris M., Abawajy J H. A rough set approach for selecting clustering attribute. Knowledge Based Systems 23 (pp. 220-231) 2010.

Herawan T., Ghazali R., Yanto I., Deris M. Rough set approach for categorical data clustering. International Journal of Database Theory and Application 3(1) (pp. 33-52) 2010.

Huang Z. Extensions to the k-means algorithm for clustering large data sets with categorical values. Data Mining and Knowledge Discovery 2(3) (pp. 283–304) 1998.

Kim D., Lee K., Lee D. Fuzzy clustering of categorical data using fuzzy centroids. Pattern Recognition Letters 25(11) (pp. 1263–1271) 2004.

Mathieu R., Gibson J. A methodology for large scale R&D planning based on cluster analysis. IEEE Transactions on Engineering Management 40(3) (pp. 283–292) 1993.

Parmar D., Wu T., Blackhurst J. MMR: An algorithm for clustering categorical data using rough set theory. Data and Knowledge Engineering 63 (pp. 879–893) 2007.

Pawlak Z., Skowron A. Rudiments of rough sets. Information Sciences 177(1) (pp. 3–27) 2007.
Pawlak Z. Rough sets: Theoretical aspects of reasoning about data. Kluwer Academic Publishers, Boston, MA 1991.
Roiger R J., Geatz M W. Data mining: A tutorial-based primer. Pearson Education 2003.
Tripathy B K., Ghosh A. A SDR: An algorithm for clustering categorical data using rough set theory. Private communication at the International IEEE Conference held in Kerala 2011.
Tripathy B K., Ghosh A. SSDR: An algorithm for clustering categorical data using rough set theory. Advances in Applied Science Research 2(3) (pp. 320–324) 2011.
Tripathy B K., Prakash M S. Kumar Ch, MMeR: An algorithm for clustering heterogeneous data using rough set theory. International Journal of Rapid Manufacturing 1(2) (pp. 189-207) 2009.
Wong K., Feng D., Meikle S., Fulham M. Segmentation of dynamic PET images using cluster analysis, IEEE Transactions on Nuclear Science 49(1) (pp. 200–207) 2002.Wu S., Liew A., Yan H., Yang M. Cluster analysis of gene expression data based on self-splitting and merging competitive learning. IEEE Transactions on Information Technology in BioMedicine 8(1) (pp. 5–15) 2004.