Journal of Computer Science and Cybernetics, V.39, N.3 (2023), 261–280
DOI no 10.15625/1813-9663/18568

A NEW INFORMATION THEORY BASED ALGORITHM FOR CLUSTERING CATEGORICAL DATA

DO SI TRUONG*, LAM THANH HIEN, NGUYEN THANH TUNG

Lac Hong University, 10 Huynh Van Nghe Street, Buu Long Ward, Bien Hoa City, Dong Nai Province, Viet Nam

Abstract. Clustering is an important technique in data mining and in machine learning. Given a set of objects, the main goal of clustering is to group objects into clusters such that objects within a cluster have high similarity to one another, but objects in different clusters have high dissimilarity. In recent years, problems of clustering categorical data have attracted much attention from the data mining research community. Several rough-set based algorithms for clustering categorical data have been proposed. These algorithms make important contributions to the problem of clustering categorical data, some of them can handle uncertainty during the clustering process, while others allow users to obtain stable results. However, they have some limitations such as they often have low accuracy and high computational complexity. In this paper, we review two baseline algorithms for use with categorical data, namely Min-Min Roughness (MMR) and Mean Gain Ratio (MGR), and propose a new algorithm, called Minimum Mean Normalized Variation of Information (MMNVI). MMNVI algorithm uses the Mean Normalized Variation of Information of one attribute concerning another for finding the best clustering attribute, and the entropy of equivalence classes generated by the selected clustering attribute for binary splitting the clustering dataset. Experimental results on real datasets from UCI indicate that the MMNVI algorithm can be used successfully in clustering categorical data. It produces better or equivalent clustering results than the baseline algorithms.

Keywords. Data mining, clustering, categorical data, information system, normalized variation of information.

1. INTRODUCTION

Clustering is a fundamental technique in data mining and machine learning. Let $D = \{x_1, x_2, \ldots, x_n\}$ 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 [1]. Clustering problem appears in many different domains such as pattern recognition, information retrieval, computer vision, bioinformatics, medicine, etc. At present, there exists a large number of clustering algorithms

*Corresponding author.
E-mail addresses: truongds@lhu.edu.vn (D.S.Truong); lthien@lhu.edu.vn (L.T.Hien); nttung@lhu.edu.vn (N.T.Tung).

in the literature. The choice of clustering algorithm depends on the data available and the purpose of the application [1, 2].

The major clustering methods can be classified as partitional, hierarchical, density-based, grid-based, and model-based clustering [1–4]. Among these, partitional and hierarchical clustering are the most popular. Partitional methods construct a single partition of dataset D into k clusters optimizing a criterion function, where k is an input parameter. Hierarchical clustering methods create a hierarchical decomposition of dataset D and this hierarchical decomposition is represented by a dendrogram. Hierarchical clustering methods can be agglomerative or divisive. Agglomerative hierarchical clustering methods start with each object in a separate group. These groups are combined successively based on a distance measure until only one group or a specific stopping criterion is achieved. Divisive hierarchical clustering methods initialize a cluster system as a single cluster of all points and gradually divide them into smaller clusters based on distance measurements. Unlike partitioning methods, hierarchical methods do not require the number of clusters, k, as an input parameter. However, a termination condition must be specified indicating when the merging or division will end [3].

Most of the earlier works on clustering have 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 consist of categorical attributes (such as gender, nationality, color, etc.) on which there is no inherent distance measure between categorical objects. Clustering categorical data is more challenging than clustering numerical data [2, 5] and clustering algorithms developed for numerical data cannot be used to cluster categorical data.

In recent years, clustering categorical data has attracted much attention from the data mining research community. Several algorithms for clustering categorical data have been proposed. The initial important algorithms include those of Huang [5, 6], Ganti et al., [7], Gibson et al., [8], and Guha et al., [9]. These algorithms make important contributions to the problem of clustering categorical data, but they are not designed handling uncertainty in the clustering process.

When studying the problem of categorical data clustering, special attention should be paid to finding techniques that allow to handle uncertainty and ambiguity, because in many real-world applications, there is no sharp boundary between clusters. Huang [6] and Kim et al. [10] have applied fuzzy set theory to clustering categorical data. However, these fuzzy set application techniques require multiple scans of the dataset to obtain the necessary stability for the membership fuzzy control parameter.

Rough set theory proposed by Z. Pawlak in the early 1980s [11], and is a relatively new soft computing tool for dealing with uncertain data. One of the major advantages of rough set theory is that it does not require any additional information about the data such as apriori probability distribution in statistics and membership function in fuzzy set theory [12]. In recent years, several rough set based divisive hierarchical clustering algorithms have been proposed for clustering categorical data [13–26].

The main idea of using rough set theory in clustering categorical data is to select a series of clustering attributes, where one of them is selected and used to split the objects at each time until all objects are clustered. Thus, the primary important task for this approach is to select from many candidates in a dataset one attribute that can best partition the objects.

The first attempt at a rough set-based technique to select clustering attributes is proposed by Mazlack et al. [17]. In [17], Mazlack et al. 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. [19] proposed the min–min–roughness (MMR) algorithm which is a "purity" rough set-based divisive hierarchical clustering algorithm for categorical data. The MMR algorithm determines the clustering attribute by the Min–Roughness (MR) criterion. In [27], Herawan et al. proposed a technique, called maximum dependency attributes (MDA), to select attributes used for divisive hierarchical clustering. MDA is constructed based on the dependency measure in rough set theory and uses it to evaluate the dependency of one attribute on the other attributes in a dataset. The MDA technique can be used to select attributes for spitting a cluster, however, it is not a completely divisive hierarchical clustering algorithm. Following the work of Herawan et al., some other researchers have also proposed new methods to select clustering attributes [14–16, 22, 23, 25, 28]. However, they have not presented specific clustering algorithms nor have they evaluated the practical effectiveness of their categorical data clustering techniques.

Qin et al. [20] proposed an information-theory-based divisive hierarchical clustering algorithm for categorical data that is implemented by selecting a clustering attribute using the mean gain ratio (MGR) and then selecting an equivalence class generated by the clustering attribute as one cluster using cluster entropy.

Recently, Wei et al. [26] systematically analyzed existing rough set-based hierarchical clustering algorithms for categorical data and introduced a uniform framework. According to this framework, a clustering algorithm is an iterative process, and each iteration comprises three main steps: (1) Select some attributes for splitting the clustering dataset; (2) Based on these selected attributes, generate bipartitions of the clustering dataset; (3) Determine which of the resulting leaf nodes should be further split. In the first step, informative attributes are selected to generate candidate bipartitions of the clustering dataset. In the second step, appropriate bipartitions are selected from the candidate bipartitions using an evaluation method. Application of the first and second steps produces a bipartition of the clustering dataset, so any given number of clusters can be reached by recursively running a divisive bisecting clustering procedure. In the third step, one of the two datasets resulting from the bipartition is selected for further splitting in the next iteration.

Although, rough set theory-based proposed categorical clustering algorithms make important contributions to the issue of clustering categorical data, they have some limitations such as they often have low accuracy and high computational complexity. Especially, on some datasets they fail or hardly select their best clustering attributes [25, 26, 29].

In this paper, we revisit two baseline divisive hierarchical clustering algorithms for use with categorical data and propose a new algorithm, called Minimum Mean Normalized Variation of Information (MMNVI). Two baseline algorithms are Min-Min Roughness (MMR) and Mean Gain Ratio (MGR). MMNVI iteratively performs only two steps on the current clustering dataset: (1) Selecting a clustering attribute; (2) Selecting an equivalence class generated by the selected clustering attribute as one cluster, and taking the union of other equivalence classes as the new clustering dataset. To implement the first step, MMNVI uses the concept of normalized variation of information in information theory which is a universal metric in the space of attributes. To perform the second step, MMNVI uses the concept

of cluster entropy. Experimental results on eight benchmark data sets from UCI show that MMNVI is a stable clustering algorithm and produces better or equivalent clustering results than the baseline algorithms.

The structure of the remainder of this paper is as follows. Section 2 presents some basic notions of rough set theory, and related concepts from information theory, and revisits two baseline algorithms. Section 3 introduces the MMNVI algorithm followed by examples for illustrative purposes. Section 4 presents our experimental results. Section 5 concludes the paper and identifies future research directions.

2. PRELIMINARIES

2.1. Some concepts of rough set theory

Definition 2.1. [11] An information system is a pair $IS = (U, A)$, where U is a non-empty finite set of objects, A is a nonempty finite set of attributes, and for every $a \in A$ there is a mapping $a : U \to V_a$, where V_a denotes the domain of a.

In the rest of this article, unless otherwise stated, we assume that all attributes in a given information system are categorical, i.e., that they have a finite and unordered domain.

In an information system $IS = (U, A)$, if an attribute is interpreted as the result of a classification, then this information system is called a decision table $DT = (U, C \cup \{d\})$, where $C \cup \{d\} = A$, $d \notin C$, C is called the condition attribute set, while d is called the decision attribute.

Definition 2.2. [11] Let $IS = (U, A)$ be an information system, $B \subseteq A$. Two elements $x, y \in U$ are said to be B-indiscernible in S if and only if $a(x) = a(y)$ for every $a \in B$.

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

Definition 2.3. [11] Let $IS = (U, A)$ be an information system, where $B \subseteq A$ 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 as follows

$$\underline{B}(X) = \{x \in U \mid [x]_B \subseteq X\} \text{ and } \overline{B}(X) = \{x \in U \mid [x]_B \cap X \neq \emptyset\}. \tag{2.1}$$

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. 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 2.4. [11] Let $IS = (U, A)$ be an information system, where $B \subseteq A$ and $X \subseteq U$. The roughness of X with respect to B is defined as

$$R_B(X) = 1 - \frac{|\underline{B}(X)|}{|\overline{B}(X)|}. \tag{2.2}$$

Obviously, $0 \leq R_B(X) \leq 1$. If $R_B(X) = 0$, X is crisp with respect to B, in other words, X is definable with respect to B. If $0 < R_B(X) \leq 1$, X is rough with respect to B, that is, B is vague with respect to X.

Definition 2.5. [11] Let $IS = (U, A)$ 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 $P \Longrightarrow_k Q$, if

$$k = \gamma_P(Q) = \frac{\sum_{X \in Q} |\underline{P}(X)|}{|U|}. \tag{2.3}$$

2.2. Related concepts from information theory

The main aim of our work is to introduce an algorithm for clustering categorical data. So, we need some special measurements to measure the disorder (uncertainty) in the column vector associated with an attribute and the intra-class similarity of a cluster. Such measurements would be the entropy of an attribute and the entropy of a cluster presented below.

Given information system $IS = (U, A)$, and attribute $a \in A$. The information system IS can be viewed as a statistical population and attribute a as a discrete random variable. Suppose $V_a = \{x_1, x_2, \ldots, x_h\}$, $U/\{a\} = \{X_1, X_2, \ldots, X_h\}$. Then the probability distribution of a can be determined by

$$P(a = x_i) = p(x_i) \ = |X_i|/|U|, \quad i = 1, \ldots, h. \tag{2.4}$$

Other related probability distributions can be similarly defined. Suppose

$$V_a = \{x_1, x_2, \ldots, x_h\},\ U/\{a\} = \{X_1, X_2, \ldots, X_h\},$$

$$V_b = \{y_1, y, \ldots, y_g\},\ U/\{b\} = \{Y_1, Y_2, \ldots, Y_g\},$$

then the joint probability distribution $P(a, b)$ of a and b, and the conditional probability distribution $P(a|b)$ of a given b are defined respectively as following

$$\begin{gathered} P(a = x_i, b = y_j) = p(x_i, y_j) = |X_i \cap Y_j|/|U|, \\ P(X = x_i \mid Y = y_j) = p(x_i \mid y_j) = |X_i \cap Y_j|/|Y_j|, \\ i = 1, \ldots, m,\ j = 1, \ldots, n. \end{gathered} \tag{2.5}$$

Definition 2.6. [29, 30] Let $IS = ((U, A))$ be an information system, attribute $a \in A$. Shannon's entropy (entropy for short) of a is defined by the following expression [29]

$$H(a) = -\sum_{i=1}^{h} p(x_i) \log_2 p(x_i), \tag{2.6}$$

and by the convention $0\log_2 0 = 0$.

For an attribute a, entropy $H(a)$ is a metric that measures the degree of disorder (uncertainty) in the column vector associated with attribute a. The smallest possible value for entropy is 0, which occurs when all components in the associated vector are the same. In other words, there is no disorder in the vector. The maximum value of entropy is $\log |V_a|$, which occurs when all components are different. The larger the value of entropy, the more disorder there is.

Definition 2.7. [17,30] Let $S = (U, A)$ be an information system, where $A = \{a_1, a_2, \ldots, a_p\}$. Assuming the attributes in A are independent of each other, we define the entropy of dataset $X \subseteq U$ as follows

$$\text{Entropy}\,(X) = H_X\,(a_1) + H_X\,(a_2) + \cdots + H_X\,(a_p)\,, \tag{2.7}$$

where $H_X\,(a_i)$ denotes the entropy of attribute a_i on X and is calculated by (2.6), $i = 1, \ldots, p$.

The smaller the entropy of X, the more similar the objects in X are. Therefore, the entropy of a cluster has been used by many authors as a measure to determine the intra-class similarity of a cluster [18, 20, 31, 32].

Definition 2.8. [29, 30] Let $S = (U, A)$ be an information system, $a, b \in A$. The joint entropy $H\,(a, b)$ of a and b is defined by

$$H\,(a, b) = -\sum_{i=1}^{h}\sum_{j=1}^{g} p\,(x_i, y_j) \log_2 p\,(x_i, y_j), \tag{2.8}$$

where $p\,(x_i, y_j) = |X_i \cap Y_j|/|U|$, $i = 1, 2, \ldots, h$ and $j = 1, 2, ..., g$.

The join entropy $H\,(a, b)$ is the measure of the amount of uncertainty which two attributes a and b contain.

Definition 2.9. [29,30] Let $S = (U, A)$ be an information system, $a, b \in A$. The conditional entropy of a with respect to b denoted by $H\,(a \mid b)$ is defined as

$$H\,(a \mid b) = -\sum_{j=1}^{g} p\,(y_j) \sum_{i=1}^{h} p\,(x_i \mid y_j) \log_2 p\,(x_i \mid y_j) = \sum_{j=1}^{g} p\,(y_j)\, H\left(a \,\middle|\, b = y_j\right). \tag{2.9}$$

The conditional entropy $H\,(a \mid b)$ quantifies the uncertainty of a random variable a when the outcome of another random variable b is known. We also have $H\,(a \mid b) = H\,(a, b) - H(b)$.

Definition 2.10. [29,30] Let $S = (U, A)$ be an information system. The mutual information between the two attributes $a, b \in A$ is defined as

$$I\,(a; b) = H\,(a) - H\,(a \mid b) = H\,(b) - H\,(b \mid a) = \; H\,(a) + \; H\,(b) - H\,(a, b)\,. \tag{2.10}$$

Mutual information $I\,(a; b)$ is non-negative and symmetric, i.e., $I\,(a; b) \geq 0$ and $I\,(a; b) = I(b; a)$. It measures the information that a and b share; It tells us how much the knowledge of one of the two attributes reduces uncertainty about the other one. Mutual information between a and b is also known as information gain of a with respect to b.

2.3. MMR algorithm

The MMR algorithm was proposed by Parmar et al. in [19]. It is one of the most successful and pioneering rough set-based hierarchical clustering algorithms for categorical data [20, 26]. MMR algorithm determines the clustering attribute using roughness measure (Definition 2.4), which allows it to have the ability to deal with uncertainty.

Given the actual clustering dataset (CDataset) and the set of attributes A, the MMR algorithm performs three main steps in each iteration of the clustering process as follows:

(1) For each attribute $a_i \in A$, let the partition of the actual clustering dataset generated by a_i be $CDataset/a_i = \{X_1, X_2, \ldots, X_g\}$, calculate $MR(a_i)$, the minimum roughness value of a_i with respect to each $a_j \in A$, $a_j \neq a_i$, using formula

$$\mathrm{MR}\,(a_i) = \min_{(a_j \in A) \wedge (j \neq i)} \Big(\mathrm{Rough}_{a_j}\,(a_i)\Big), \tag{2.11}$$

where $\mathrm{Rough}_{a_j}\,(a_i) = \sum_{k=1}^{g} R_{a_j}(X_k)/g$, and $R_{a_j}(X_k)$ is the roughness of X_k with respect to a_j, and is calculated by formula (2.2). After that, it determines the clustering attribute a^*, such that $a^* = \mathrm{argmin}_{a_i \in A}\,\{MR\,(a_i)\}$.

(2) Select the splitting equivalence class $X_0 \in Dataset/a^*$ satisfying

$$X_0 = \arg\min_{X_k} \left(\sum_{a_j \in A,\ a_j \neq a^*} R_{a_j}(X_k) \right),$$

and set $X_0' = CDataset - X_0$.

(3) Among X_0 and X_0', choose the set which has the bigger number of objects as the new clustering dataset, and output the remaining set as a cluster.

The iterative clustering process continues until the number of clusters obtained equals the pre-defined number k of clusters.

MMR is considered one of the most successful rough set-based clustering algorithms. Besides the ability to handle uncertainty in the clustering process, the MMR algorithm is a powerful clustering algorithm and it is capable of handling large data sets. Although MMR still has two major drawbacks: (1) MMR tends to choose the clustering attribute with fewer values [20, 27], so if an attribute has only a single value, it will be selected, resulting in the termination of clustering. (2) The MMR algorithm chooses the dataset with more objects for further split, which is not always consistent with the natural distribution of clusters, thus possibly generating undesirable clustering results.

2.4. MGR algorithm

The MGR algorithm was proposed by Qin et al. in [20]. It is an information theory based divisive hierarchical clustering algorithm for categorical data. MGR algorithm determines the clustering attribute using the information gain ratio which also allows MGR to have the ability to deal with uncertainty.

Given the actual clustering dataset ($CDataset$) and the set of attributes A, the MGR algorithm performs three main steps in each iteration of the clustering process as follows:

(1) For each attribute $a_i \in A$, let the partition of actual clustering dataset grnerated by a_i be $CDataset/a_i = \{X_1, X_2, \ldots, X_g\}$, calculate the mean information gain ratio (MGR) of attribute a_i, with respect to each $a_j \in A$, $a_j \neq a_i$, using the following formula

$$MGR\,(a_i) = \frac{1}{|A| - 1} \sum_{j=1, j \neq i}^{|A|} GR_{a_j}\,(a_i), \tag{2.12}$$

where $GR_{a_j}(a_i)$ is gain ratio of a_i with respect to a_j and is calculated by the following formula

$$GR_{a_j}(a_i) = \frac{I(a_i; a_j)}{H(a_i)}. \tag{2.13}$$

After that, the algorithm determines the clustering attribute a^* satisfying that

$$a^* = \arg\max_{a_j \in A} \{MGR(a_j)\}.$$

(2) Selects the splitting equivalence class $X_0 \in Dataset/a^*$ satisfying

$$X_0 = \arg\min_{X_k \in CDataset/a^*} (\text{Entropy}(X_k)).$$

(3) Output X_0 as one cluster, and take set $CDataset = CDataset - X_0$ as a new clustering dataset for the next iteration.

MGR repeats the above steps on the new clustering dataset until the number of clusters obtained equals the pre-defined number k of clusters.

In comparison with MMR, the MGR algorithm has two advantages [20]: (1) MGR does not tend to choose the clustering attribute with fewer values; (2) MGR outputs the found cluster in each iteration based on the intra-class similarity of a cluster, regardless of its cardinality and performs further decomposition on the remaining objects, which is more natural than MMR algorithm.

According to Wei et al. [26], the main disadvantage of the MGR algorithm is that it would be not suitable to use MGR to select the baseline attribute for splitting a clustering dataset. From formula (2.13), one can see that the gain ratio $GR_{a_j}(a_i)$ of one attribute a_i with respect to another attribute a_j can be very large if both the entropy of a_i and the information gain $I(a_i; a_j)$ are low. In other words, the similarity between two attributes a_i and a_j can be low even if the gain ratio $GR_{a_j}(a_i)$ is large. In such cases, it is not suitable to use MGR to select baseline attributes for splitting a cluster.

3. PROPOSED ALGORITHM

Considering the advantages and disadvantages of the two baseline algorithms above, this section introduces a new algorithm for clustering categorical data, which is called Minimum Mean Normalized Variation of Information (MMNVI).

3.1. Basic definitions and idea of MMNVI

As seen in Section 2, in a categorical information system $S = (U, A)$, each attribute in A defines a partition on the set U of objects. A good clustering of the objects should share as much information as possible with the partitions defined by each attribute in A [18, 20, 32]. Also, the smaller the entropy of the data set is the more similar the objects in it are. Keeping this in mind, in our iterative clustering algorithm, at each iteration, we expect to choose the clustering attribute whose partition is closest to those defined by other attributes. Then, in the partition defined by the chosen clustering attribute, select the equivalence class with the highest intra-class similarity as a cluster, and take the rest of the objects to form a new

clustering dataset. The above two steps will be repeated on the new clustering dataset until the number of clusters obtained equals the pre-defined number k of clusters.

To measure the distance between two attributes, the NVI (normalized variation of information) metric is used and it is defined below.

Definition 3.1. [29,30] Given an information system $IS = (U, A)$, $a_i, a_j \in A$. The normalized variation of information between a_i and a_j is defined by

$$NVI(a_i, a_j) = 1 - \frac{I(a_i; a_j)}{H(a_i, a_j)}. \tag{3.1}$$

$NVI(X, Y)$ is a metric on the space of attributes, that is, for any attributes a_i, a_j, and a_k, it satisfies:

1) $NVI(a_i, a_j) \geq 0$ and the equality holds iff $a_i = a_j$,
2) $NVI(a_i, a_j) = NVI(a_j, a_i)$,
3) $NVI(a_i, a_j) + NVI(a_j, a_k) \geq NVI(a_i, a_k)$.

Values of $NVI(a_i, a_j)$ are in the range [0,1]. Note that, $NVI(a, b)$ is a universal metric in the sense that if any other distance measure puts a and b close to each other, the NVI will also evaluate them to be close [29].

Definition 3.2. Given an information system $IS = (U, A)$, $a_i \in A$. The mean normalized variation of information of a_i with respect to each $a_j \in A$, $a_j \neq a_i$, is defined by

$$MNVI(a_i) = \frac{1}{|A| - 1} \sum_{j=1,\ j \neq i}^{|A|} NVI(a_i, a_j). \tag{3.2}$$

3.2. MMNVI algorithm

With the above idea and definitions, our MMNVI iterative algorithm works as follows.

On the first iteration, it takes the set of all objects U as the clustering dataset and performs the following three main steps:

1. Remove all the single-valued attributes.
2. Selects a clustering attribute as the one that has the smallest MNVI value.
3. Output a cluster the equivalence class generated by the partitioning attribute which has the lowest entropy, and takes as a new clustering dataset the union of other equivalence classes.

The above clustering process repeats until the number of the leaf nodes equals the pre-defined number k of clusters. Fig.1 shows the pseudo-code of the MMNVI algorithm.

We have two following remarks about the MMNVI algorithm.

1. In Step 4, if many attributes have the same smallest MNVI value, we choose the first of them.

2. In Steps 5 and 6, after having an attribute for splitting the clustering dataset, MMNVI selects the equivalence class with the lowest entropy as one cluster and takes the union of other equivalence classes as the new clustering dataset. This is because the entropy of the second dataset is larger as indicated by the following proposition. In addition, if there are multiple equivalence classes with the same lowest entropy, we select the equivalence class with the largest number of objects.

Algorithm MMNVI
Input: Set of all objects U, set of all attributes A, the given number of clusters k.
Output: Clustering of U
Begin
Step 1: Set current number of cluster $CNC = 1$;
Set $CDataset = U$ // $CDataset$ denotes the actual clustering data set
Step 2: $B = A$;
for each $a_i \in B$
determine partition $CDataset/Ind\{a_i\}$;
if $|CDataset/Ind\{a_i\}| = 1$ // all objects in $CDataset$ have the same values of a_i
$B = B$ - a_i // exclude attribute a_i;
endif
endfor
Step 3: **for** each $a_i \in B$
calculate $MNVI(a_i)$ using formula (3.2);
endfor
Step 4: Determine the clustering attribute a^* which satisfies
$a^* = \text{argmin}_{a_i \in B} MNVI(a_i)$;
Step 5: Determine partition $CDataset/Ind\{a^*\} = \{X_1, X_2, \ldots, X_h\}$;
$X = \text{argmin}_{X_i}(ent(X_i))$ for $X_i \in CDataset/Ind\{a^*\}$;
Step 6: Output X as one cluster;
$CNC = CNC + 1$;
if $CNC < k$
$CDataset = CDataset - X$;
go to Step 2;
else
output $CDataset$ as the last cluster;
endif
End

Proposition 1. *Let $S = (U, A)$ be a information system, attribute $a \in A$, and*

$$U/a = \{X_1, X_2, \ldots, X_h\}.$$

If

$$\text{Entropy}(X_1) = \min\{\text{Entropy}(X_1), \text{Entropy}(X_2), \ldots, \text{Entropy}(X_h)\},$$

then

$$\text{Entropy}(X_2 \cup X_3 \cup \cdots \cup X_h) \geq \text{Entropy}(X_1).$$

Proof. Let $U' = X_2 \cup X_3 \cup \cdots \cup X_h$. Because, U/a is the partition of U generated by a, every object in X_i has the same value x_i on attribute a. For each attribute $b \in A - \{a\}$, we have

$$H_{U'}(b \mid a) = \sum_{i=2}^{h} \frac{|X_i|}{|U'|} H_{U'}(b \mid a = x_i) = \sum_{i=2}^{h} \frac{|X_i|}{|U'|} H_{X_i}(b).$$

On the other hand, according to formula (2.10), $H_{U'}(b) \geq H_{U'}(b \mid a)$. Therefore,

$$H_{U'}(b) \geq \sum_{i=2}^{h} \frac{|X_i|}{|U'|} H_{X_i}(b),$$

$$\sum_{b\in A-\{a\}} H_{U'}(b) \geq \sum_{b\in A-\{a\}} \sum_{i=2}^{h} \frac{|X_i|}{|U'|} H_{X_i}(b) = \sum_{i=1}^{h} \frac{|X_i|}{|U'|} \sum_{b\in A-\{a\}} H_{X_i}(b).$$

Note that, $H_{X_i}(a) = 0$ for each $i = 2, 3, \ldots, h$, we have

$$\sum_{b\in A-\{a\}} H_{X_i}(b) = \sum_{b\in A} H_{X_i}(b),$$

$$\sum_{b\in A} H_{U'}(b) \geq \sum_{i=1}^{h} \frac{|X_i|}{|U'|} \sum_{b\in A} H_{X_i}(b) = \sum_{i=1}^{h} \frac{|X_i|}{|U'|} \text{Entropy}(X_i) \geq \text{Entropy}(X_1),$$

or

$$\text{Entropy}(X_2 \cup X_3 \cup \cdots \cup X_h) \geq \text{Entropy}(X_1).$$

■

3.3. Example

Let's consider the categorical information system given in Table 1, there are eight objects with seven categorical attributes. Suppose we want to split this set of objects into 3 clusters. We have

$$U = \{1, 2, \ldots, 8\},$$

$$A = \{\text{Degree}, \text{English}, \text{Experience}, \text{IT}, \text{Mathematics}, \text{Programming}, \text{Statistics}\}.$$

Table 1: An information system of student enrollment qualification in [27]

Student	Degree $a1$	English $a2$	Experience $a3$	It $a4$	Mathematics $a5$	Programming $a6$	Statistics $a7$
1	Ph.D	Good	Medium	Good	Good	Good	Good
2	Ph.D	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

At the first iteration, MMNVI takes the set of all eight objects U as the clustering dataset. Because all seven attributes are multivalued attributes, no attribute is removed. MMNVI determines the best clustering attribute for the first binary split.

The normalized variation of information of $a1$ with respect to $a2$, for example, is calculated by (3.1) as follows

$$U/Ind(\{a1\}) = \{X_1, X_2, X_3\} = \{\{1, 2\}, \{3, 4, 5, 6\}, \{7, 8\}\},$$

$$U/Ind(\{a2\}) = \{Y_1, Y_2, Y_3\} = \{\{1\}, \{2, 3, 4, 5, 6, 7\}, \{8\}\},$$

$$U/Ind(\{a1, a2\}) = \{Y_1, Y_2, Y_3\} = \{\{1\}, \{2\}, \{3, 4, 5, 6\}, \{7\}, \{8\}\},$$

$$H(a1) = 1.5, H(a2) = 1.0613, \ H(a1, a2) = 2, I(a1; a2) = 0.5613.$$

So, $NVI(a1, a2) = 1 - \ I(a1; a2)/Ha1, a2) = 0.7194$.

Following similar process, we get the normalized variation of information of $a1$ with respect to $a3$, $a4$, $a5$, $a6$, and $a7$. These values are 0.4591, 0.3333, 0.7500, 0.7500, and 0.8403.

Table 2 lists the normalized variation of information of each attribute with respect to each other attribute.

Table 2: Mean certainty of each attribute with respect to each other attribute

Attributes	*NVI* (Normalized Variation of Information)							MNVI
	$a1$	$a2$	$a3$	$a4$	$a5$	$a6$	$a7$	
$a1$		0.7194	0.4591	0.3333	0.7500	0.7500	0.8403	**0.642**
$a2$	0.7194		0.7910	0.8221	0.8620	0.8620	0.8221	0.8131
$a3$	0.4591	0.7910		0.7925	0.7925	0.7925	1.0000	0.7713
$a4$	0.3333	0.8221	0.7925		1.0000	1.0000	0.8958	0.8073
$a5$	0.7500	0.8620	0.7925	1.0000		0.0000	0.8958	0.7167
$a6$	0.7500	0.8620	0.7925	1.0000	0.0000		0.8958	0.7167
$a7$	0.8403	0.8221	1.0000	0.8958	0.8958	0.8958		0.8916

The last column of Table 2 lists the Mean Normalized Variation of Information of each ai with respect to each $aj \in A$, $aj \neq ai$.

After having the Mean Normalized Variation of Information of each attribute, the best clustering attribute for the first binary split is determined. The attribute $a1$ has the minimum MNVI value. Thus, $a1$ is selected as the clustering attribute, and binary splitting is conducted. We have

$$U/Ind\left(\{a1\}\right) = \{X_1, X_2, X_3\} = \{\{1,2\}, \{3,4,5,6\}, \{7,8\}\},$$

$$\text{Entropy}\left(X_1\right) = 1,\ \text{Entropy}\left(X_2\right) = 2.8113,\ \text{Entropy}\left(X_3\right) = 2.$$

Because X_1 has the smallest entropy value, it is chosen to form a cluster. The set of remaining objects is $Z_1 = \{3,4,5,6,7,8\}$ and Z_1 will be the new clustering dataset for iteration 2.

In iteration 2, there are three attributes which give us the minimum value of the Mean Normalized Variation of Information. These attributes are 1, 3 and 4. Choosing attribute 3 as the clustering attribute, we have $Z_1/a3 = \{Y_1, Y_2\} = \{\{3,4,5,6\},\{7,8\}\}$. Because $\text{Entropy}\left(Y_1\right) = 2.8113 > \text{Entropy}\left(Y_2\right) = 2$, Y_2 is selected as the second cluster.

With the pre-defined number of clusters $k = 3$, MMNVI gives us three clusters $C_1 = \{1,2\}$, $C_2 = \{7,8\}$, and $C_3 = \{3,4,5,6\}$.

3.4. Computational complexity of MMNVI algorithm

Now, let's consider the time complexity of the MMNVI algorithm. Suppose that in the dataset for clustering, there are n objects, m attributes, and k is the pre-defined number of clusters. To split the dataset into k clusters, the algorithm has to run $k-1$ iterations. In each iteration, the MMNVI algorithm needs to calculate the MNVI values of all m attributes. To compute the MNVI value of an arbitrary attribute a_i, the time to determine equivalence classes is n, and the time to calculate NVI with respect to other attributes is $n(m-1)$. Thus, the time to calculate the values of MNVI for all m attributes is $m(n + n\,(m-1)) = nm^2$. In addition, the MMNVI algorithm needs to calculate $n(k-1)$ times the entropy of the

equivalence classes. The time to calculate the entropy of an equivalence class is no more than m. Thus, the time to be spent in calculating entropy is much less than $(k-1)nm$. To sum up, the expected time complexity of the MMNVI algorithm is polynomial, which is $\mathrm{O}(knm + knm^2)$.

According to references [26] and [28], the time complexity of the MMR and MGR algorithms is the same and is $\mathrm{O}(knm + km^2l)$.

3.5. Theoretical advantage of MMNVI algorithm

Similar to MMR and MGR, our MMNVI algorithm is capable of handling the uncertainty in the clustering process, (by using information measures to measure the uncertainty of an object set), needs k the pre-defined number of clusters as the input parameter, does not depend on initial values and the input order of data, and can output a stable clustering result (repeated runs produce the same result).

Compared with the MMR and MGR algorithms, the MMNVI algorithm has three main improvements:

(1) In each iteration, before selecting the clustering attribute, MMNVI removes all the single-valued attributes, thus, MMNVI can avoid premature stopping of the clustering process.

(2) The MMNVI algorithm evaluates a candidate attribute by the Mean Normalized Variation of Information rather than by the Min–Min-Roughness because the rationality of the partition induced by an attribute should be reflected on all of the attributes instead of on just one best attribute.

(3) The MMNVI algorithm takes the dataset with a larger entropy for further splitting (clustering), which helps to improve the clustering accuracy. This is superior to the MMR algorithm, which selects the dataset with more objects.

4. EXPERIMENTAL RESULTS

To test MMNVI, an implementation system was developed in R programming language, and tested on real-life data sets. Besides the MMNVI algorithm, we also repeat two baseline algorithms, MMR and MGR, to compare with MMNVI.

4.1. Benchmark datasets

To evaluate the clustering performance of MMNVI, we used 8 real-life data sets obtained from the UCI Machine Learning Repository [34], including Soybean small, Zoo, Votes, Breast Cancer Wisconsin, Mushroom, Balance Scale, Car Evaluation, and Chess. The information about the data sets is given in Table 3.

In the Breast Cancer data set, there are 16 objects with missing values. In our implementation, these objects were deleted, so the number of left objects was 683.

Three algorithms are sequentially run on all data sets. Each algorithm requires the number of clusters to be clustered as an input parameter. In our experiments, the number of clusters was set to be the known number of class labels. For instance, the number of clusters was set to 4 for the Car evalution data set.

Table 3: Eight UCI data sets

Datasets name	Abbreviations	Number of objects	Number of attributes	Number of classes
Soybean small	Soybean	47	359	4
Breast Cancer Wisconsin	Breast	683 (699)	9	2
Car Evalution	Car	1728	6	4
Congressional Voting Records	Vote	435	16	2
Chess	Chess	3196	36	2
Mushroom	Mushroom	8124	22	2
Balance scale	Balance	625	4	3
Zoo	Zoo	101	16	7

4.2. Performance evaluation methods

To evaluate the performance of clustering algorithms, we use three widely used indexes. These indexes are Overal Purity, Adjusted Rand Index (ARI), and Normalized Mutual Information (NMI).

Table 4: Contingency table

$\Omega\backslash C$	c_1	c_2	…	c_J	*sums*
ω_1	n_{11}	n_{12}	…	n_{1J}	a_1
ω_2	n_{21}	n_{23}	…	n_{2J}	a_2
…	…	…	…	…	…
ω_I	n_{I1}	n_{I2}	…	n_{IJ}	a_I
sums	b_1	b_2	…	b_J	$\sum_{ij} n_{ij} = n$

Suppose that data set has n objects, $\Omega = \{\omega_1, \omega_2, \ldots, \omega_I\}$ and $C = \{c_1, c_2, \ldots, c_J\}$ represent the clustering result and the original classification, respectively, cluster ω_i has a_i objects, class c_j has b_j objects, and n_{ij} is number of objects that are in both cluster ω_i and class c_j. Thus, we have the contingency table as given in Table 4.

With the above notations, three evaluation indexes are defined as follows:

(1) The purity of a cluster ω_i and the Overall Purity of clusters are, respectively, as follows [31]

$$Purity\,(\omega_i) = \frac{1}{a_i}\max_j\,(n_{ij}) \quad \text{and} \quad \text{OP} = \frac{\sum_{i=1}^{I}\max_j\,(n_{ij})}{n}. \tag{4.1}$$

The OP lies between 0 and 1. According to this measure, a higher value of OP indicates a better clustering result. With perfect clustering, i.e., clustering contains only pure clusters, the overall purity gives a value of 1.

(2) The Adjusted Rand Index (ARI) can be expressed as [31]

$$ARI\,(\Omega, C) = \frac{\sum_{i,j} C_{n_{ij}}^2 - \left[\sum_i C_{a_i}^2 \sum_j C_{b_j}^2\right]/C_n^2}{\frac{1}{2}\left[\sum_i C_{a_i}^2 + \sum_j C_{b_j}^2\right] - \left[\sum_i C_{a_i}^2 \sum_j C_{b_j}^2\right]/C_n^2}, \tag{4.2}$$

where $C_{a_i}^2 = a_i(a_i-1)/2$, $C_{b_j}^2 = b_j(b_j-1)/2$, $C_{n_{ij}}^2 = n_{ij}(n_{ij}-1)/2$.

The ARI lies between 0 and 1. When the clustering result and the original classification agree perfectly, the adjusted rand index is 1.

(3) The Normalized Mutual Information (NMI) is built upon fundamental concepts from information theory. Given two clusterings Ω and C, their entropies, joint entropy, and mutual information are defined naturally via the marginal and joint distributions of data items in Ω and C, respectively, as follows [30,31]

$$H(\Omega) = -\sum_{i=1}^{I} \frac{a_i}{n} \log_2 \frac{a_i}{n}, \tag{4.3}$$

$$H(\Omega, C) = -\sum_{i=1}^{I} \sum_{j=1}^{J} \frac{n_{ij}}{n} \log_2 \frac{n_{ij}}{n}, \tag{4.4}$$

$$I(\Omega;C) = H(\Omega) + H(C) - H(\Omega, C). \tag{4.5}$$

Then, normalized mutual information (NMI) is calculated by

$$\text{NMI} = \frac{2I(\Omega;C)}{H(\Omega) + H(\text{C})}. \tag{4.6}$$

NMI lies in [0,1], equal to 1 when the two clusterings are identical, and 0 when they are independent, that is, sharing no information about each other.

4.3. Clustering results

We first present the detailed clustering results of MMNVI, then we show the Overall Purity, Adjusted Rand Index, and Normalized Mutual Information values obtained by MMR, MGR, and MMNVI algorithms on 8 data sets.

For the Soybean small data set, as the known number of class labels in this data set is 4, therefore, the number of clusters is set to 4 for all three MMNVI, MMR and MGR algorithms in the test. The clustering distributions on the Soybean data set are summarized in Table 5. It is evident from Table 5 that the MMNVI algorithm obtains clustering results with 47 objects belonging to the majority class label. Thus, the Overall Purity of the clusters is 1. The Adjusted Rand and the Normalized Mutual Information are equal to 0.4601, and 0.6511, respectively.

Table 5: Results of MMNVI on the Soybean Small data set

Cluster	Class 1	Class 2	Class 3	Class 4	Cluster purity	Overall purity	ARI	NMI
1	10	0	0	0	1	1	0.4601	0.6511
2	10	0	0	0	1			
3	10	0	0	0	1			
4	0	0	0	17	1			

Tables 6 – 12 show the clustering results of the MMNVI algorithm on seven other UCI data sets.

Table 6: Results of MMNVI on the Breast Cancer Wisconsin data set

Cluster	Class 1	Class 2	Cluster purity	Overall purity	ARI	NMI
1	369	4	0.9893	0.8843	0.59	0.5446
2	75	235	0.7581			

Table 7: Results of MMNVI on the Car Evaluation data set

Cluster	Class 1	Class 2	Class 3	Class 4	Cluster purity	Overall purity	ARI	NMI
1	108	0	324	0	0.75	0.7384	0.0071	0.0452
2	115	23	268	26	0.6204			
3	108	0	324	0	0.75			
4	72	0	360	0	0.8333			

Table 8: Results of MMNVI on the Votes data set

Cluster	Class 1	Class 2	Cluster purity	Overall purity	ARI	NMI
1	200	8	0.9615	0.8276	0.4279	0.4009
2	67	160	0.7048			

Table 9: Results of MMNVI on the Chess data set

Cluster	Class 1	Class 2	Cluster purity	Overall purity	ARI	NMI
1	922	895	0.5074	0.5307	0.0036	0.0034
2	605	774	0.5613			

Table 10: Results of MMNVI on the Mushroom data set

Cluster	Class 1	Class 2	Cluster purity	Overall purity	ARI	NMI
1	0	36	1	0.5224	-0.0011	0.009
2	4208	3880	0.5203			

Table 11: Results of MMNVI on the Zoo data set

Cluster	Class 1	Class 2	Class 3	Class 4	Class 5	Class 6	Class 7	Cluster purity	Overall purity	ARI	NMI
1	41	0	0	0	0	0	0	1	0.7327	0.3211	0.3707
2	1	0	0	0	0	0	0	1			
3	9	6	0	2	1	2	0	0.45			
4	11	3	0	4	2	1	1	0.5			
5	3	0	0	3	0	0	0	0.5			
6	1	0	0	0	0	0	0	1			
7	0	0	0	0	0	8	2	0.8			

Table 12: Results of MMNVI on the Balance Scale data set

Cluster	Class 1	Class 2	Class 3	Cluster purity	Overall purity	ARI	NMI
1	10	17	98	0.784	0.6784	0.1234	0.1346
2	10	17	98	0.784			
3	28	228	119	0.608			

4.4. Comparison with MMR and MGR algorithms

With the same process MMR and MGR are applied to the 8 real-life data sets. The Overall Purity values of the three algorithms are summarized in Table 13.

Table 13: Overall Purity values of three algorithms on 8 data sets

Algorithms	Soybean	Breast	Car	Votes	Chess	Mushroom	Balance	Zoo	Average
MMR	0.8298	0.6559	0.7002	0.6138	0.5225	**0.7002**	0.6352	0.9109	0.6961
MGR	**1**	0.5	0.6998	0.5	**0.5338**	0.6775	0.6352	**0.9307**	0.6846
MMNVI	**1**	**0.8843**	**0.7384**	**0.8276**	0.5307	0.5224	**0.6784**	0.7327	**0.7393**

Out of 8 data sets, MMNVI has the highest OP on the five data sets, specifically on the Soybean Small, Breast Cancer Wisconsin, Car evalution, and Votes and Balance scale.

MMR has the highest OP on the Mushroom. MGR has the highest OP on the Soybean Small, Chess, and Zoo. The last column of Table 13 shows the average OP of each algorithm on 8 data sets. On average, MMNVI achieves the highest Overall Purity.

Table 14: ARI values of three algorithms on 8 data sets

Algorithms	Soybean	Breast	Car	Votes	Chess	Mushroom	Balance	Zoo	Average
MMR	**0.6738**	0.0101	**0.0129**	-0.0068	0.0004	0.0129	0.1011	0.913	0.2146
MGR	0.4601	0.1465	**0.0129**	0.106	0.0036	**0.1254**	0.1011	**0.9617**	0.2397
MMNVI	0.4601	**0.59**	0.0071	**0.4279**	**0.0036**	-0.0011	**0.1234**	0.3211	**0.2415**

The ARI values of the three algorithms are summarized in Table 14. From this table, we see that MMNVI also has the highest ARI value on the four data sets, Breast Cancer Wisconsin, Votes, Chess, and Balance scale. MMR has the lowest ARI value across all 8 data sets. MGR has the highest ARI value on Car evalution, Mushroom, and Zoo. The last column of Table 14 shows the average ARI of each algorithm on 8 data sets. On average, the MMNVI algorithm also achieves the highest ARI value.

The NMI values of the three algorithms are summarized in Table 15. MMNVI has the highest NMI value on the three data sets, Breast Cancer Wisconsin, Votes, and Balance scale. MMR has the highest NMI on the Soybean Small, Car evalution, and Mushroom. MGR has the highest NMI on Chess and Zoo. The last column of Table 15 shows the average NMI of each algorithm on 8 data sets.

Table 15: NMI values of three algorithms on 8 data sets

Algorithms	Soybean	Breast	Car	Votes	Chess	Mushroom	Balance	Zoo	Average
MMR	**0.8264**	0.0405	**0.0621**	0.0041	0.0052	**0.0621**	0.0902	0.913	0.2504
MGR	0.6511	0.5445	0.0481	0.401	**0.017**	0.0246	0.1344	**0.9617**	**0.3478**
MMNVI	0.6511	**0.5446**	0.0452	**0.4009**	0.0034	0.009	**0.1346**	0.3707	0.2699

An important observation is that MMNVI does much better than other algorithms on the Breast Cancer, Votes, and Balance scale. Breast Cancer and Votes are data sets which have a balanced class distribution. MGR does much better than other algorithms on the Zoo.

In summary, MMNVI is a stable clustering algorithm and produces better or equivalent clustering results than the MMR and MGR algorithms.

5. CONCLUSION

In this paper, we have proposed a new divisive hierarchical clustering algorithm, called MMNVI (Minimum Mean Normalized Variation of Information), for categorical data. MMNVI algorithm uses the Mean Normalized Variation of Information of one attribute with respect to another attribute for finding the best clustering attribute, and the entropy of equivalence classes generated by the selected clustering attribute for binary splitting the clustering dataset. MMNVI is easy to install. Experimental results on real-life data sets from UCI indicate that the MMNVI algorithm can be used successfully in clustering categorical data. It is a stable clustering algorithm and produces better or equivalent clustering

results than the baseline algorithms. It can be applied to the data sets which have balanced class distribution, such as Breast Cancer and Votes.

For future work, we will attempt to: (1) Enable the MMNVI algorithm to automatically discover the number of clusters. Instead of specifying the number of clusters, we can let the MMNVI algorithm stop splitting the clustering data set when the entropies of all the leaf nodes are lower than a predefined threshold value; (2) Extend MMNVI to handle both numerical and categorical data; (3) Perform more experiments on larger data sets with more objects and more attributes.

REFERENCES

[1] J. Han and M. Kamber, *Data Mining: Concepts and Techniques*, 3rd ed. Morgan Kaufmann Publishers, 2012.

[2] S. Mesakar and M. S. Chaudhari, "Review paper on data clustering of categorical data," *International Journal of Engineering Research & Technology*, vol. 1, no. 10, December 2012.

[3] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: A review," *ACM Computing Surveys*, vol. 31, no. 3, pp. 264–323, 1999.

[4] D. Jyot, "Clustering categorical data using rough sets: a review," *International Journal of Advanced Research in IT and Engineering*, vol. 2, no. 12, pp. 30–37, 2013.

[5] Z. Huang, "Extensions to the k-averages algorithm for clustering large data sets with categorical values," *Data Mining and Knowledge Discovery*, vol. 2, no. 3, pp. 283–304, 1998.

[6] Z. Huang and M. K. Ng, "A fuzzy k-modes algorithm for clustering categorical data," *IEEE Transactions on Fuzzy Systems*, vol. 7, no. 4, pp. 446–452, 1999.

[7] V. Ganti, J. Gehrke, and R. Ramakrishnan, "Cactus–clustering categorical data using summaries," in *Proceeding of Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1999, pp. 73–83.

[8] D. Gibson, J. Kleinberg, and P. Raghavan, "Clustering categorical data: An approach based on dynamical systems," *Very Large Data Bases J.*, vol. 8, no. 3–4, pp. 222–236, 2000.

[9] S. Guha, R. Rastogi, and K. Shim, "Rock: A robust clustering algorithm for categorical attributes," in *Proceeding of 15th ICDE*, 1999, pp. 512–521.

[10] D. Kim, K. Lee, and D. Lee, "Fuzzy clustering of categorical data using fuzzy centroids," *Pattern Recognition Letters*, vol. 25, no. 1, pp. 1263–1271, 2004.

[11] Z. Z. Pawlak, *Rough Sets - Theoretical Aspects of Reasoning about Data.* Dordrecht: Kluwer Academic Publishers, 1991.

[12] A. Skowron and S. Dutta, "Rough sets: Past, present, and future," *Natural Computing*, vol. 17, no. 4, pp. 855–876, 2018.

[13] M. M. Baroud, S. Z. M. Hashim, J. U. Ahsan, and A. Zainal, "Positive region: An enhancement of partitioning attribute based rough set for categorical data," *Periodicals of Engineering and Natural Sciences*, vol. 8, no. 4, pp. 2424–2439, December 2020.

[14] W. A. Hassanein, "Clustering algorithms for categorical data using concepts of significance and dependence of attributes," *European Scientific Journal*, vol. 10, no. 3, pp. 381–400, 2014.

[15] W. Hassanein and A. Elmelegy, "An algorithm for selecting clustering attribute using significance of attributes," *International Journal of Database Theory and Application*, vol. 6, no. 5, pp. 53–66, 2013.

[16] P. Kumar and B. Tripathy, "Mmer: An algorithm for clustering heterogeneous data using rough set theory," *International Journal of Rapid Manufacturing*, vol. 1, no. 2, pp. 189–207, 2009.

[17] L. J. Mazlack, A. He, Y. Zhu, and S. Coppock, "A rough set approach in choosing clustering attributes," in *Proceedings of the ISCA 13th International Conference (CAINE 2000)*, 2000, pp. 1–6.

[18] I.-K. Park and G.-S. Choi, "Rough set approach for clustering categorical data using information-theoretic dependency measure," *Information Systems*, vol. 4, pp. 289–295, 2015.

[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, pp. 879–893, 2007.

[20] H. Qin, X. Ma, T. Herawan, and J. M. Zain, "Mgr: An information theory based hierarchical divisive clustering algorithm for categorical data," *Knowledge-Based Systems*, vol. 67, pp. 401–411, 2014.

[21] G. K. Singh and S. Mandal, "Cluster analysis using rough set theory," *Journal of Informatics and Mathematical Sciences*, vol. 9, no. 3, pp. 509–520, 2017.

[22] B. Tripathy and A. Ghosh, "Sdr: An algorithm for clustering categorical data using rough set theory," in *Recent Advances in Intelligent Computational Systems, IEEE*, 2011, pp. 867–872.

[23] B. K. Tripathy, A. Goyal, R. Chowdhury, and P. A. Sourav, "Mmemer: An algorithm for clustering heterogeneous data using rough set theory," *I.J. Intelligent Systems and Applications*, vol. 8, pp. 25–33, 2017.

[24] J. Uddin, R. Ghazali, and M. M. Deris, "An empirical analysis of rough set categorical clustering techniques," *PLOS ONE*, vol. 12, no. 1, 2017.

[25] J. Uddin, R. Ghazali, J. H. Abawajy, H. Shah, N. A. Husaini, and A. Zeb, "Rough set based information theoretic approach for clustering uncertain categorical data," *PLOS ONE*, May 13 2022. [Online]. Available: https://doi.org/10.1371/journal.pone.0265190

[26] W. Wei, J. Liang, X. Guo, P. Song, and Y. Sun, "Hierarchical division clustering framework for categorical data," *Neurocomputing*, vol. 341, pp. 118–134, 2019.

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

[28] P. C. Xuyen, D. S. Truong, and N. T. Tung, "An information-theoretic metric based method for selecting clustering attribute," in *Proceedings of 9th National Conference on Fundamental and Applied Information Technology*, 2016, pp. 31–40.

[29] Y. Y. Yao, "Information-theoretic measures for knowledge discovery and data mining," in *Entropy Measures, Maximum Entropy Principle and Emerging Applications. Studies in Fuzziness and Soft Computing*, Karmeshu, Ed. Springer, Berlin, Heidelberg, 2003, vol. 119. [Online]. Available: https://doi.org/10.1007/978-3-540-36212-8_6

[30] F. M. Reza, *An Introduction to Information Theory.* New York: Dover Publications, 1994.

[31] M. Halkidi, Y. Batistakis, and M. Vazirgiannis, "On clustering validation techniques," *Journal of Intelligent Information Systems*, vol. 17, pp. 107–145, 2001.

[32] J. McCaffrey, "Data clustering using entropy minimization," 2018. [Online]. Available: http://visualstudiomagazine.com/Articles/2013/02/01/Data-Clustering-Using-Entropy-Minimization.aspx?Page=2&p=1

Received on July 21, 2023
Accepted on August 27, 2023