Knowledge and Information Systems (2025) 67:4113–4138
https://doi.org/10.1007/s10115-024-02317-x

REGULAR PAPER

Clusterability test for categorical data

Lianyu Hu[1] · Junjie Dong[1] · Mudi Jiang[1] · Yan Liu[2] · Zengyou He[1,3]

Received: 2 August 2024 / Revised: 1 November 2024 / Accepted: 17 December 2024 /
Published online: 6 January 2025

Abstract
The objective of clusterability evaluation is to check whether a clustering structure exists within the data set. As a crucial yet often-overlooked issue in cluster analysis, it is essential to conduct such a test before applying any clustering algorithm. If a data set is unclusterable, any subsequent clustering analysis would not yield valid results. Despite its importance, the majority of existing studies focus on numerical data, leaving the clusterability evaluation issue for categorical data as an open problem. Here, we present TestCat, a testing-based approach to assess the clusterability of categorical data in terms of an analytical p-value. The key idea underlying TestCat is that clusterable categorical data possess many strongly associated attribute pairs and hence the sum of Chi-squared statistics of all attribute pairs is employed as the test statistic for p-value calculation. We apply our method to a set of benchmark categorical data sets, showing that TestCat outperforms those solutions based on existing clusterability evaluation methods for numeric data. To the best of our knowledge, our work provides the first way to effectively recognize the clusterability of categorical data in a statistically sound manner.

Keywords Clusterability · Categorical data · Clustering structure · Statistical hypothesis test

✉ Zengyou He
zyhe@dlut.edu.cn

Lianyu Hu
hly4ml@gmail.com

Junjie Dong
jd445@qq.com

Mudi Jiang
792145962@qq.com

Yan Liu
yliu0414@qq.com

1 School of Software, Dalian University of Technology, Dalian 116620, China

2 School of Software Engineering, Dalian University, Dalian 116622, China

3 Key Laboratory for Ubiquitous Network and Service Software of Liaoning Province, Dalian 116620, China

1 Introduction

Clustering, as a multivariate analysis tool, is widely used in diverse fields such as biology and social sciences, aiming to identify a collection of homogeneous groups within the data. In the absence of a unified definition for natural clusters [1–4], various properties [5] that affect clustering quality often exhibit conflicting behavior [6], posing challenges for users when assessing the effectiveness of clustering methods. Running different clustering algorithms on identical input data can yield considerably distinct results. Typically, these approaches are evaluated within the context of specific applications [7], and a proper algorithm is selected by utilizing validity indices [8] on the produced clusters. It is essential to recognize that both clustering algorithms and validity metrics hinge on the presence of underlying clustering structure. Otherwise, the functions built upon that may not output meaningful results, and utilizing these outcomes could lead to untrustworthy conclusions.

To determine whether data are clusterable before employing any clustering algorithms, different methods have been suggested to assess *clusterability* [9–12], which aim to evaluate the extent of clustering structure present within the data. Despite the inherent ambiguity of clustering, it is evident that a data set generated from a single Gaussian distribution is unclusterable, as any partitioning on it would be spurious. Given this observation, current solutions [11, 13] predominantly employ a hypothesis testing approach, specifically the multimodality tests in one dimension (Dip test [14] and Silverman test [15]). As the resulting p-value provides a probabilistic interpretation, one can determine whether the underlying structure is significant by using a significance threshold. For multivariate data sets, dimensionality reduction techniques, like PCA (principal component analysis), are necessary to project the data into one-dimensional space, enabling the application of multimodality tests. In addition, several intuitive visual tools, such as VAT (visual assessment of tendency) [16] and dissimilarity plots [17], can help reveal the clustering structure in data sets.

Categorical variables are pervasive in real-world data sets [18] and serve to represent qualitative information from a variety of fields, like marital status in sociological questionnaires [19] and cell types in biology [20]. However, all the aforementioned clusterability measures are designed specifically for numerical variables, as the dimensionality reduction techniques and the hypothesis testing procedures are not suitable for handling categorical variables. For example, in a single categorical attribute, the discrete nature of attribute values in distinct groupings, making it inappropriate to assume that the random variable follows a unimodal distribution. Furthermore, common distance metrics that several clusterability evaluation methods rely on, cannot be directly applied to this type of data due to the lack of geometric properties. Consequently, in order to employ these conventional methods, one must convert categorical data into one-dimensional numerical values by utilizing dissimilarity measures explicitly designed for categorical data [21], which are practically ineffective (see results in 4.4). Hence, identifying the existence of a categorical clustering structure remains an overlooked open challenge. When designing a clusterability evaluation method for categorical data, it is necessary to employ new significance testing strategies that differ from what have been deployed for numerical data.

Here, we develop a testing-based approach, TestCat, to assess the clusterability of categorical data. Intuitively, in a clusterable categorical data set, samples in each cluster should be quite similar to each other. Accordingly, many attribute values within the cluster will be positively associated with each other. Furthermore, over-expressed attribute values in one cluster will be negatively associated with those over-expressed ones in another cluster. Hence, it can be expected that there will be many associated attribute pairs in a clusterable categorical data

set. Based on the above observations, we employ the Chi-squared test to measure the association for each pair of attributes and utilize the sum of Chi-squared statistics of all attribute pairs as the test statistic for clusterability validation.[1] By imposing an independence assumption on attribute pairs, we can derive a p-value to determine the clusterability of any categorical data set. Validation studies conducted on real data sets have demonstrated the effectiveness and efficiency of our method. More importantly, the empirical results have shown that commonly used methods designed for numerical data may fail to distinguish between clusterable and unclusterable categorical data sets.

In summary, the main contributions of this work are as follows:

- The problem of clusterability assessment for categorical data is conceptualized as an issue of testing association among attributes. To the best of our knowledge, this is the first attempt to develop a method specifically designed for assessing the clusterability of categorical data.
- A p-value based on Chi-squared tests is derived as a validation index to determine the clusterability of categorical data. Experimental results on real categorical data sets demonstrate that the proposed method is both effective and robust, while also being comparable in efficiency to other methods.
- To date, how to empirically compare the performance of different clusterability evaluation methods for categorical data is still an open problem. A feasible and reasonable pipeline is presented and adopted in our empirical studies, which may serve as a practical validation strategy for future research toward this direction.

The rest of this paper is organized as follows: Sect. 2 provides a review of existing methods that are closely related to our topic. Section 3 offers a detailed description of our method. Section 4 presents the results, including experiments on real data sets and further analysis. Lastly, Sect. 5 concludes the paper.

2 Related work

Currently, there are no clusterability evaluation methods explicitly designed for categorical data. Consequently, we will provide separate overviews of clustering methods for categorical data and clusterability evaluation methods for numerical data. It is important to note that clustering algorithms for categorical data do not readily lend themselves to clusterability evaluation methods, and at the same time, the testing-based approaches used in existing clusterability evaluation methods are not well-suited for handling categorical data.

2.1 Categorical data clustering

Existing categorical data clustering algorithms [22], which include various approaches such as partition-based, model-based, density-based, and linkage-based hierarchical methods, can generally be classified into two main types based on the objective functions they employ: k-means-type [23] and others [24, 25]. The k-means-type algorithms quantify each cluster by calculating the sum of the squared distances [26] between each object and its cluster center (i.e., mode) [27]. Additionally, some of these algorithms utilize categorical-to-numerical techniques [28, 29], transforming categorical data into numerical data, and then applying

[1] Notably, TestCat focuses on pairwise attribute associations and may not capture complex cluster structures determined by higher-order interactions among multiple attributes where pairwise associations are absent.

the original k-means algorithm. In contrast, other algorithms use measures like entropy [30, 31] or category utility [32, 33] to quantify a cluster. They assess the difference between the observed category distribution in a cluster and the expected distribution under a random assignment of objects to clusters.

The value of objective function is data-dependent and can vary significantly across different data sets. The numerical value itself does not offer insights into the quality of individual clusters or the overall partition. Instead, it represents a local optimum among a myriad of possible partitions. Furthermore, even when we have the global optimum values, they do not signify the presence of a clustering structure or indicate whether the clusters are obtained from a randomized data set. Notably, methods based on category utility merely measure the difference between the observed category distribution and the expected category distribution, failing to provide a significance-based score in terms of p-value. Anyway, all these objective functions can be used to evaluate the goodness of clustering results; however, they cannot be directly used for the purpose of clusterability evaluation.

2.2 Clusterability evaluation methods

Without relying on any specific partitioning, standard-compliant clusterability evaluation methods have been proposed, as seen in [11, 13]. Meanwhile, algorithm-dependent methods, such as those presented in [34, 35], have been excluded from consideration as they do not adequately meet the practical requirements of clusterability assessment. More specifically, the output values of the clusterability evaluation function can be used to declare data as either clusterable or unclusterable. This goal is well achieved by testing-based approaches that employ a p-value as the validation index, providing a statistically meaningful alternative to other methods.

For numerical data, testing-based clusterability evaluation methods conduct the analysis on either the original data or transformed data. A single cluster typically arises from a homogeneous Gaussian distribution, so the presence of multiple clusters suggests a deviation from this pattern. Accordingly, detecting multimodality in a data set with statistical methods such as Dip [14] and Silverman [15] serves as a proxy for clusterability assessment. Similarly, the concept that a single cluster corresponds to spatial randomness has led to the development of the Hopkins statistic [36] and the PHI statistic [9].

When addressing categorical data, the approach to clusterability evaluation needs to be tailored to accommodate the unique characteristics of this data type. The test statistics mentioned earlier are not directly applicable to the original data, as categories cannot be treated as numerical values. Likewise, dimensionality reduction techniques typically used for numerical data, such as PCA, are inappropriate. Instead, the analysis should start with distance measures crafted for categorical variables [21]. Following this step, dimensionality reduction methods compatible with distance values, such as multidimensional scaling (MDS) [37], t-distributed stochastic neighbor embedding (tSNE) [38] and uniform manifold approximation and projection (UMAP) [39], can be utilized.

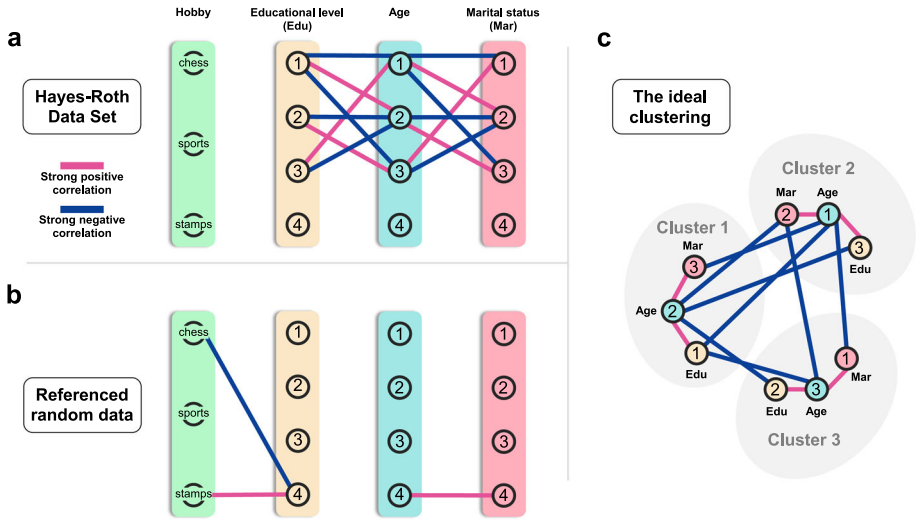

Fig. 1 Parallel coordinate plots are utilized to display the strong association among neighboring attribute pairs in both the **a** Hayes-Roth data set and its **b** referenced random data. The attribute values are represented by "chess", "sports", "stamps" or "1", "2", "3", "4". The strength of association is determined by standardized residuals (refer to Supplementary Method 1). A standardized residual value exceeding 2 signifies a strong positive association, while a value less than −2 indicates a strong negative association. **c** The ideal clustering extracted from Hayes-Roth data set contains attributes with strong positive associations within each cluster, while those with strong negative associations are distributed across separate clusters. Each of these clusters is represented by specific categories (color figure online)

3 Methods

3.1 Overview of TestCat

TestCat quantifies the clusterability by measuring the strength of association between all pairs of attributes in a given categorical data set. We will demonstrate that a marked difference in association strength exists between clusterable and non-clusterable (random) data. For instance, in the Hayes-Roth data set, exemplars can be grouped into three distinct club memberships, which plays a role in facilitating psychological experiments [40]. These qualitative attributes are hobby, educational level, age, and marital status. A conspicuous phenomenon arises from the data (Fig. 1): the attribute values of latter three attributes display an abundance of both positive and negative associations in the original data set (Fig. 1a). Given this strongly associated categories across different attributes, we postulate that the inherent clustering structure (groups) is derived from these distinct attribute combinations as shown in Fig. 1c. Contrarily, within a random data (generated as explained in Sect. 3.3), the values of hobby attribute (a distractor feature) may exhibit a few spurious associated relationship with one attribute value of educational level, while there is only one positive association among the remaining attribute values. We can thus conclude that the original data set is composed of many associated attribute pairs, whereas artificial random data sets lacking a significant clustering structure do not present a comparable magnitude of these associations.

To evaluate the overall strength of association among all attribute pairs and derive a clusterability index in terms of p-value, the TestCat approach consists of the following

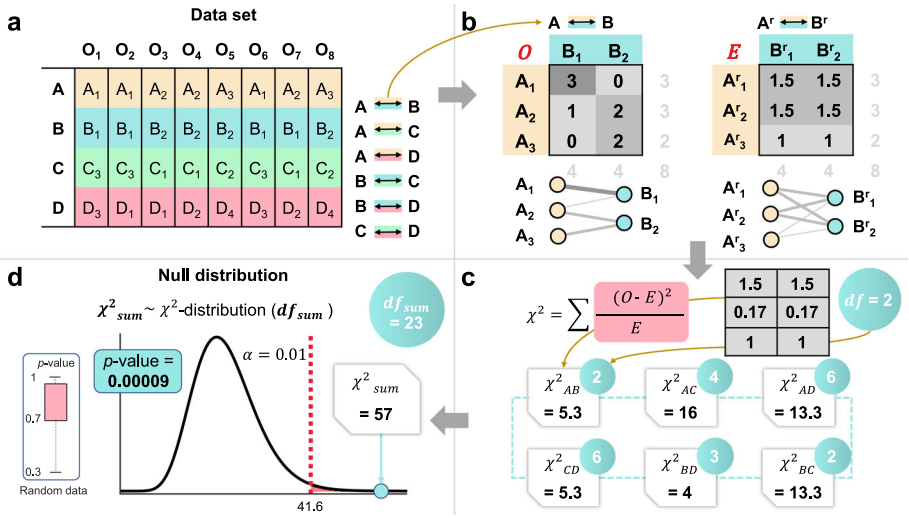

Fig. 2 The TestCat method conducts clusterability analysis on a given categorical data set by providing a p-value. **a** A toy categorical data set of four attributes ("A", "B", "C", "D") yields six different attribute pairs that need to be tested. **b** For each attribute pair, such as "A-B", both its observed (O) and expected (E) 3×2 contingency tables are constructed and used to calculate a Chi-squared test statistic. The darker cells or lines indicate a higher frequency of co-occurrence between the two attribute values in the data set. **c** The Chi-squared test statistics for all six attribute pairs, along with their respective degrees of freedom (df), are collected and employed to derive a final p-value. **d** Under an imposed assumption (refer to Methods in 3.2), a single p-value can be calculated from the sum of test statistics and its corresponding null distribution. A boxplot of the p-values obtained from referenced random data sets is also displayed (color figure online)

steps (Fig. 2). We initiate our analysis by constructing contingency tables for each pair of categorical variables/ attributes (e.g., "A" and "B" shown in Fig. 2b) of a given data set. Under the assumption of independence between two attributes, the table for the expected frequencies of attribute value pairs can be obtained from (light-colored) marginal frequencies. We employ the Chi-squared test to evaluate the association between each pair of attributes. Then, we collect all Chi-squared test statistics and their corresponding degrees of freedom for each attribute pair (Fig. 2c). Finally, we use the sum of Chi-squared test statistics as final test statistic, which will still follows a Chi-squared distribution under the assumption that all variables for attribute pairs are independent (Fig. 2d).

The significance testing issue underlying our TestCat method can be formally formulated as follows. Let S denote the number of all attribute pairs. For the s-th attribute pair, H_{0s} denotes the null hypothesis that these two attributes are independent and H_{1s} denotes the alternative hypothesis that they are not independent. The global null hypothesis is a composite of the individual null hypotheses H_{0s} for all S tests, represented as:

$$H_0 : \bigcap_{s=1}^{S} H_{0s},$$

contrasted against the global alternative hypothesis:

$$H_1 : \bigcup_{s=1}^{S} H_{1s}, \tag{1}$$

where the global alternative hypothesis implies that at least one H_{0s} is false. In the TestCat framework, the global null hypothesis H_0 posits that each pair of attributes exhibits no association, where the s-th individual null hypothesis H_{0s} aligns with the Chi-squared test affirming the independence of the s-th attribute pair. Conversely, the global alternative hypothesis H_1 indicates the presence of an association between at least one pair of attributes. Furthermore, associations across multiple attribute pairs can cumulatively strengthen H_1. To determine whether H_0 should be rejected, it is necessary to combine all p-values from S individual Chi-squared tests into a single composite p-value.

Various methods exist for combining p-values [41], such as Fisher's method and Stouffer's method. These approaches create a new test statistic by integrating all p-values, which is distinct from the statistic employed in each individual test. In our TestCat approach, we opt for simplicity by using the sum of Chi-squared statistics of all attribute pairs as the new test statistic, which also follows a Chi-squared distribution under the independence assumption on different pairs of attributes.

3.2 Chi-squared test

In the Chi-squared test used in our approach, a Chi-squared statistic is obtained for each attribute pair. Consider the s-th attribute pair $\langle A, B\rangle_s$, where $A = [A_1, A_2, \cdots, A_{Q_a^{(s)}}]$ and $B = [B_1, B_2, \cdots, B_{Q_b^{(s)}}]$ are two distinct attributes, containing $Q_a^{(s)}$ and $Q_b^{(s)}$ categories, respectively. The observed frequencies of attribute value pairs are presented in a contingency table as follows:

$$O^{(s)} = \begin{array}{c|ccc} & \mathrm{B}_1 & \cdots & B_{Q_b^{(s)}} \\ \hline \mathrm{A}_1 & O_{11}^{(s)} & O_{1j}^{(s)} & O_{1Q_b^{(s)}}^{(s)} \\ \vdots & O_{i1}^{(s)} & O_{ij}^{(s)} & O_{iQ_b^{(s)}}^{(s)} \\ \mathrm{A}_{Q_a^{(s)}} & O_{Q_a^{(s)}1}^{(s)} & O_{Q_a^{(s)}j}^{(s)} & O_{Q_a^{(s)}Q_b^{(s)}}^{(s)} \end{array}. \tag{2}$$

The i-th row and column marginal frequencies, as well as the grand total of the contingency table, are represented as follows:

$$\begin{aligned} O_{i\cdot}^{(s)} &= \sum_{j}^{Q_b^{(s)}} O_{ij}^{(s)}, \; O_{\cdot j}^{(s)} = \sum_{i}^{Q_a^{(s)}} O_{ij}^{(s)}, \\ O_{\cdot\cdot}^{(s)} &= \sum_{i}^{Q_a^{(s)}} \sum_{j}^{Q_b^{(s)}} O_{ij}^{(s)}. \end{aligned} \tag{3}$$

Then, the expected frequencies can be calculated and represented as follows:

$$E_{ij}^{(s)} = (O_{i\cdot}^{(s)} \cdot O_{\cdot j}^{(s)})/O_{\cdot\cdot}^{(s)},$$

$$E^{(s)} = \begin{array}{c|ccc} & \mathrm{B}_1 & \cdots & B_{Q_b^{(s)}} \\ \hline \mathrm{A}_1 & E_{11}^{(s)} & E_{1j}^{(s)} & E_{1Q_b^{(s)}}^{(s)} \\ \vdots & E_{i1}^{(s)} & E_{ij}^{(s)} & E_{iQ_b^{(s)}}^{(s)} \\ \mathrm{A}_{Q_a^{(s)}} & E_{Q_a^{(s)}1}^{(s)} & E_{Q_a^{(s)}j}^{(s)} & E_{Q_a^{(s)}Q_b^{(s)}}^{(s)} \end{array} . \tag{4}$$

For each attribute pair, we calculate the Chi-squared statistic, $\chi^2_{(s)}$, based on $O_{ij}^{(s)}$ and $E_{ij}^{(s)}$. These individual Chi-squared statistics are then summed to yield a collective measure for all attribute pairs. The formula is as follows:

$$\chi^2_{sum} = \sum_{s}^{S} \chi^2_{(s)} = \sum_{s}^{S} \sum_{i}^{Q_a^{(s)}} \sum_{j}^{Q_b^{(s)}} \frac{(O_{ij}^{(s)} - E_{ij}^{(s)})^2}{E_{ij}^{(s)}}, \tag{5}$$

where $S = \binom{M}{2}$ denotes the total number of all unique attribute pairs and M is the number of attributes of the given data set. For the s-th Chi-squared statistic, the degrees of freedom $df^{(s)}$ are equivalent to the number of frequencies in the corresponding contingency table. The aggregate degrees of freedom for the summed Chi-squared statistic χ_{sum} are calculated as follows:

$$df_{sum} = \sum_{s}^{S} df^{(s)} = \sum_{s}^{S} (Q_a^{(s)} - 1) \cdot (Q_b^{(s)} - 1) . \tag{6}$$

Finally, a p-value is derived from the summed Chi-squared statistic and its aggregate degrees of freedom by utilizing the Chi-squared cumulative distribution function. Note that the sum of Chi-squared random variables follows a Chi-squared distribution if these variables are independent [18]. Here, we impose an assumption on the independence among variables for attribute pairs in order to obtain an analytical p-value for clusterability evaluation.

3.3 Randomized data generation

For the m-th attribute of randomized data, we assume that both the m-th attribute of original data set (ODS) and the m-th attribute of its corresponding randomized data set (CRDS) follow the same categorical distribution, which corresponds to the marginal distribution of the contingency table.

To generate a CRDS, we successively generate M random permutations of data objects in ODS. According to the m-th permutation, we specify the attribute values for the m-th attribute for each data object in CRDS. More precisely, the corresponding attribute value for the j-th object in the CRDS is specified to be one that are taken by the j-th object of ODS in the permutation.

3.4 Theoretical justification

To justify our method, we need to prove that a larger test statistic (corresponding to a smaller p-value) indicates the better clusterability of the data set. However, this task is quite difficult in a general setting because (1) the test statistic is complicated since it is the sum of multiple Chi-squared statistics, and (2) there is still no consensus on the definition of clusterability of a data set. Nevertheless, we try to provide a theoretical justification on a special case that the

data set is only composed of two categorical attributes. More precisely, we first investigate the case that each attribute is composed of only two distinct attribute values. Then, we extend the analysis to the general case that each attribute can have more than two attribute values.

3.4.1 Special case: binary categorical attributes

Suppose that these two attributes are denoted by $A = \{A_1, A_2\}$ and $B = \{B_1, B_2\}$, respectively. The corresponding contingency table is given as follows, where C_{ij} is the number of co-occurrence of attribute values A_i and B_j. Obviously, such a data set is composed of four natural clusters, where each cluster is characterized by $A = A_i$ and $B = B_j$. In other words, each cluster consists of C_{ij} identical attribute value pairs. Under this setting, it is easy to see that the distances among samples within each cluster are all zeros. Hence, to quantify the clusterability of such a data set, we only need to check the distances between samples from different clusters. More precisely, the following measure is employed to quantify the clusterability of a data set.

	B_1	B_2	
A_1	C_{11}	C_{12}	$C_{1\cdot}$
A_2	C_{21}	C_{22}	$C_{2\cdot}$
	$C_{\cdot 1}$	$C_{\cdot 2}$	$C_{\cdot\cdot}$

(7)

Definition 1 Let the data set DS be characterized by the contingency table shown in 7. The normalized separation of DS, denoted as $Sep_{norm}(DS)$, is defined as follows:

$$\mathrm{Sep}_{\mathrm{norm}}(DS) = \frac{\mathrm{Sep}(DS)}{S_{\mathrm{total}}(DS)}, \tag{8}$$

where $Sep(DS)$ is the sum of all pairwise inter-cluster Hamming distances and S_{total} is the number of all pairwise inter-cluster samples. The concise expression for $Sep(DS)$ can be calculated as follows:

$$\begin{aligned}\mathrm{Sep}(DS) &= \mathrm{Sep}(C_{11}, C_{12}) + \mathrm{Sep}(C_{11}, C_{21}) + \mathrm{Sep}(C_{11}, C_{22}) \\ &\quad + \mathrm{Sep}(C_{12}, C_{21}) + \mathrm{Sep}(C_{12}, C_{22}) + \mathrm{Sep}(C_{21}, C_{22}) \\ &= C_{11}C_{12} + C_{11}C_{21} + 2C_{11}C_{22} + 2C_{12}C_{21} + C_{12}C_{22} + C_{21}C_{22} \\ &= C_{11}(C_{\cdot\cdot} - C_{11}) + C_{22}(C_{\cdot\cdot} - C_{22}) + 2C_{12}C_{21}\end{aligned}. \tag{9}$$

Given $C_{22} = C_{\cdot 2} - (C_{1\cdot} - C_{11})$ and $C_{12}C_{21} = (C_{1\cdot} - C_{11})(C_{\cdot 1} - C_{11})$, we have $Sep(DS) = C_{11}(C_{\cdot\cdot} - C_{11}) + (C_{11} - C_{1\cdot} + C_{\cdot 2})(C_{\cdot\cdot} - C_{11} + C_{1\cdot} - C_{\cdot 2}) + 2(C_{1\cdot} - C_{11})(C_{\cdot 1} - C_{11})$. If we set $C_{11} = n$, $C_{\cdot\cdot} = N$, $C_{1\cdot} = x$, $C_{\cdot 1} = y$, $C_{\cdot 2} = z$, then we can write as:

$$\begin{aligned}\mathrm{Sep}(DS) &= n(N-n) + (n - x + z)(N - n + x - z) + 2(x - n)(y - n) \\ &= Nn - n^2 + (N - n + x - z)n - x(N - n + x - z) \\ &\quad + z(N - n + x - z) + 2xy + 2n^2 - 2(x + y)n \\ &= (2N - 2z - 2y)n + (z - x)N - x^2 - z^2 + 2x(y + z)\end{aligned}.$$

Given $y + z = C_{\cdot 1} + C_{\cdot 2} = C_{\cdot\cdot} = N$, we have

$$\begin{aligned} \mathrm{Sep}(DS) &= (z - x)N - (x^2 + z^2) + 2xN \\ &= (x + z)N - ((x + z)^2 - 2xz) \\ &= (N - x - z)(x + z) + 2xz \\ &= (C_{\cdot\cdot} - C_{1\cdot} - C_{\cdot 2})(C_{1\cdot} + C_{\cdot 2}) + 2C_{1\cdot}C_{\cdot 2} \end{aligned} \quad . \tag{10}$$

Now we turn to express another element $S_{total}(DS)$ as follows:

$$\begin{aligned} S_{\mathrm{total}}(DS) &= S(C_{11}, C_{12}) + S(C_{11}, C_{21}) + S(C_{11}, C_{22}) \\ &\quad + S(C_{12}, C_{21}) + S(C_{12}, C_{22}) + S(C_{21}, C_{22}) \\ &= C_{11}C_{12} + C_{11}C_{21} + C_{11}C_{22} + C_{12}C_{21} + C_{12}C_{22} + C_{21}C_{22} \\ &= \big(C_{11}(C_{\cdot\cdot} - C_{11}) + C_{22}(C_{\cdot\cdot} - C_{22}) + 2C_{12}C_{21}\big) \\ &\quad - (C_{12}C_{21} + C_{11}C_{22}) \\ &= \mathrm{Sep}(DS) - (C_{12}C_{21} + C_{11}C_{22}) \end{aligned} \quad . \tag{11}$$

Lemma 1 *Given a* 2×2 *contingency table with fixed marginal frequencies, if* C_{11} *exceeds* $\lambda = \frac{2C_{1\cdot}+C_{\cdot 1}-C_{\cdot 2}}{4}$*, then the* $Sep_{\mathrm{norm}}(DS)$ *is directly proportional to* C_{11}*:* $C_{11} > \lambda \rightarrow Sep_{\mathrm{norm}}(DS) \propto C_{11}$.

Proof Given that $\mathrm{Sep}_{\mathrm{norm}}(DS) = \frac{\mathrm{Sep}(DS)}{\mathrm{Sep}(DS)-(C_{12}C_{21}+C_{11}C_{22})}$ where $\mathrm{Sep}(DS)$ is a constant under fixed marginal frequencies, $\mathrm{Sep}_{\mathrm{norm}}(DS)$ is only dependent on the term $(C_{12}C_{21} + C_{11}C_{22})$. Thus, we rewrite $(C_{12}C_{21} + C_{11}C_{22})$ with C_{11} as the only variable as follows:

$$\begin{aligned} C_{12}C_{21} + C_{11}C_{22} &= C_{11}(C_{\cdot 2} - (C_{1\cdot} - C_{11})) \\ &\quad + (C_{1\cdot} - C_{11})(C_{\cdot 1} - C_{11}) \\ &= C_{\cdot 2}C_{11} - C_{1\cdot}C_{11} + (C_{11})^2 \\ &\quad + C_{1\cdot}C_{\cdot 1} - C_{1\cdot}C_{11} - C_{\cdot 1}C_{11} + (C_{11})^2 \\ &= 2(C_{11})^2 + (C_{\cdot 2} - 2C_{1\cdot} - C_{\cdot 1})C_{11} + C_{1\cdot}C_{\cdot 1} \\ &= f(C_{11}) \end{aligned} \quad . \tag{12}$$

Taking the derivative of f with respect to C_{11}, we get $f'(C_{11}) = 4C_{11} + (C_{\cdot 2} - 2C_{1\cdot} - C_{\cdot 1})$ and find the critical point: $\lambda = \frac{2C_{1\cdot}+C_{\cdot 1}-C_{\cdot 2}}{4}$.

Thus, when $C_{11} > \lambda$, $f(C_{11})$ is an increasing function. Now, since $\mathrm{Sep}_{\mathrm{norm}}(DS)$ is functionally dependent on $f(C_{11})$, it follows that the clusters become more distinct in terms of separation as C_{11} increases, thereby supporting the claim that a larger C_{11} within the 2×2 contingency table corresponds to more distinct clusters. □

Lemma 2 *Given a* 2×2 *contingency table with fixed marginal frequencies, assume without loss of generality that* C_{11} *is greater than its expected frequency* E_{11}*. Under this condition, the Chi-squared statistic* $\chi^2(DS)$ *is directly proportional to* C_{11}*: if* $C_{11} > E_{11}$*, then* $\chi^2(DS) \propto C_{11}$.

Proof Given the condition $C_{11} > E_{11} = \frac{C_{1\cdot}C_{\cdot 1}}{C_{\cdot\cdot}}$, we have $C_{\cdot\cdot}C_{11} - C_{1\cdot}C_{\cdot 1} > 0$. Furthermore, according to the definition of the Chi-squared statistic for a 2×2 contingency table, we have $\chi^2(DS) = \frac{C_{\cdot\cdot}(C_{11}C_{22}-C_{12}C_{21})^2}{C_{1\cdot}C_{\cdot 1}C_{\cdot 2}C_{2\cdot}} = C_{DS}(C_{11}C_{22} - C_{12}C_{21})^2$ where C_{DS} is a constant under

fixed marginal frequencies. The $\chi^2(DS)$ is only dependent on the squared term $(C_{11}C_{22} - C_{12}C_{21})^2$. Thus, we can express $(C_{11}C_{22} - C_{12}C_{21})$ with C_{11} as the only variable as follows:

$$\begin{aligned} C_{11}C_{22} - C_{12}C_{21} &= C_{11}(C_{\cdot 2} - (C_{1\cdot} - C_{11})) - (C_{1\cdot} - C_{11})(C_{\cdot 1} - C_{11}) \\ &= C_{\cdot 2}C_{11} - C_{1\cdot}C_{11} + (C_{11})^2 \\ &\quad - C_{1\cdot}C_{\cdot 1} + C_{1\cdot}C_{11} + C_{\cdot 1}C_{11} - (C_{11})^2 \\ &= (C_{\cdot 2} - C_{1\cdot} + C_{1\cdot} + C_{\cdot 1})C_{11} - C_{1\cdot}C_{\cdot 1} \\ &= C_{\cdot\cdot}C_{11} - C_{1\cdot}C_{\cdot 1} \\ &> 0 \end{aligned} \quad . \tag{13}$$

Now that we have established $C_{11}C_{22} - C_{12}C_{21} = C_{\cdot\cdot}C_{11} - C_{1\cdot}C_{\cdot 1}$ according to Eq. 13. Under the given condition $C_{11} > E_{11}$, $(C_{\cdot\cdot}C_{11} - C_{1\cdot}C_{\cdot 1})^2 = (C_{11}C_{22} - C_{12}C_{21})^2$ is an increasing function. This indicates a stronger association between attributes A and B as C_{11} increases, thereby supporting the claim that a larger C_{11} within the 2×2 contingency table is indicative of an increased strength of association. □

Theorem 1 *Given a* 2×2 *contingency table with fixed marginal frequencies, for any two data sets* DS_1 *and* DS_2 *sampled according to these marginal distributions, if* $\chi^2(DS_1) > \chi^2(DS_2) > \lambda^* = C_{DS}\big(C_{\cdot\cdot}\lambda - C_{1\cdot}C_{\cdot 1}\big)^2$ *where* $C_{DS} = \frac{C_{\cdot\cdot}}{C_{1\cdot}C_{\cdot 1}C_{\cdot 2}C_{2\cdot}}$ *and* $\lambda = \frac{2C_{1\cdot}+C_{\cdot 1}-C_{\cdot 2}}{4}$, *then it follows that* $Sep_{\text{norm}}(DS_1) > Sep_{\text{norm}}(DS_2)$.

Proof Let us denote p and q as the count variables in the 2×2 contingency table corresponding to the positively associated cell C_{11} for data sets DS_1 and DS_2, respectively. Since the marginal frequencies are fixed, the Chi-squared statistics $\chi^2(DS_1)$ and $\chi^2(DS_2)$ along with the same constant can be expressed in terms of p, q as $C_{DS}(C_{\cdot\cdot}p - C_{1\cdot}C_{\cdot 1})^2$ and $C_{DS}(C_{\cdot\cdot}q - C_{1\cdot}C_{\cdot 1})^2$ according to Eq. 13. Then, we can rewrite the condition $\chi^2(DS_2) > \lambda^*$ as follows:

$$\begin{aligned} \chi^2(DS_2) > \lambda^* &\Rightarrow C_{DS}(C_{\cdot\cdot}q - C_{1\cdot}C_{\cdot 1})^2 \\ &> C_{DS}(C_{\cdot\cdot}\lambda - C_{1\cdot}C_{\cdot 1})^2 \\ &\Rightarrow C_{\cdot\cdot}q - C_{1\cdot}C_{\cdot 1} \\ &> C_{\cdot\cdot}\lambda - C_{1\cdot}C_{\cdot 1} \\ &\Rightarrow q > \lambda = \frac{2C_{1\cdot} + C_{\cdot 1} - C_{\cdot 2}}{4} \end{aligned} \quad . \tag{14}$$

Combining the results from Lemmas 1 and 2 with Eq. 14, we finalize the proof as follows: Lemma 2 provides that $\chi^2(DS)$ is directly proportional to C_{11}. Consequently, given $\chi^2(DS_1) > \chi^2(DS_2) > \lambda^*$, we apply Eq. 14 to deduce that $p > q > \lambda$. Furthermore, Lemma 1 suggests that for $C_{11} > \lambda$, which is satisfied here as $p > q > \lambda$, the normalized separation measure $\text{Sep}_{\text{norm}}(DS)$ will increase. Therefore, we can assert that $\text{Sep}_{\text{norm}}(DS_1)$ is greater than $Sep_{norm}(DS_2)$. □

3.4.2 Extension to general categorical attributes

We now extend our theoretical justification to the case where the two attributes A and B are categorical variables with more than two categories. Let $A = \{A_1, A_2, \cdots, A_r\}$ and $B = \{B_1, B_2, \ldots, B_s\}$, where r and s are the numbers of distinct categories of A and B, respectively.

An important observation is that the Hamming distance between two samples computed on the original categorical variables is identical to the Hamming distance computed after

one-hot encoding these variables. Given this equivalence, we can conceptually extend our previous results from the binary case to the general case. The Chi-squared statistic for the contingency table of A and B is:

$$\chi^2(DS) = \sum_{i=1}^{r}\sum_{j=1}^{s} \frac{\left(C_{ij} - E_{ij}\right)^2}{E_{ij}}, \tag{15}$$

where C_{ij} is the observed frequency of the category pair (A_i, B_j), and $E_{ij} = \frac{C_{i\cdot}C_{\cdot j}}{N}$ is the expected frequency under independence.

A larger Chi-squared statistic indicates stronger associations between certain categories of A and B, which promotes the formation of clusters with higher intra-cluster similarity and greater inter-cluster separation. Since each summation term in the equation above can be interpreted as the case of binary categorical attributes, we can infer that this cumulative larger Chi-squared statistic corresponds to a higher normalized separation measure $\mathrm{Sep}_{\mathrm{norm}}(DS)$, indicating better clusterability of the data set.

This extension suggests that our method is applicable to attributes with more than two categories, and that the relationship between the $\chi^2(DS)$ and the $\mathrm{Sep}_{\mathrm{norm}}(DS)$ holds in the general case. However, we acknowledge that deriving an explicit analytical form in this general setting is complex and challenging, requiring more in-depth analysis and effort.

4 Results

4.1 An illustrative example for two attributes

In the context of a 2×2 contingency table, under certain mild conditions, we have demonstrated in Theorem 1 that one data set with a larger Chi-squared test statistic will exhibit a better clusterability. As an illustrative example, let us consider a series of data sets, each containing 100 students, with two attributes: "grades in math" and "grades in physics", where the co-occurrence of categories "Good-in-Math" and "Good-in-Physics" indicates a positive association, represented by the count C_{11}, consistent with the notation used in Sect. 3.4.

We set C_{11} to be 20, 15, and 10 on three different data sets, each with fixed marginal distributions $(C_{\cdot 1}, C_{\cdot 2}, C_{1\cdot}, C_{2\cdot})$. These data sets are designated as DS_1, DS_2, DS_3, respectively. The contingency tables for these data sets are as follows:

DS_1	Good-in-Physics	Poor-in-Physics	
Good-in-Math	20 (Cluster 1)	5 (Cluster 2)	25
Poor-in-Math	20 (Cluster 3)	55 (Cluster 4)	75
	40	60	100
DS_2	Good-in-Physics	Poor-in-Physics	
Good-in-Math	15 (Cluster 1)	10 (Cluster 2)	25
Poor-in-Math	25 (Cluster 3)	50 (Cluster 4)	75
	40	60	100
DS_3	Good-in-Physics	Poor-in-Physics	
Good-in-Math	10 (Cluster 1)	15 (Cluster 2)	25
Poor-in-Math	30 (Cluster 3)	45 (Cluster 4)	75
	40	60	100

From the above contingency tables, we can calculate their corresponding χ^2, p-value and Sep_{norm} as follows:

Measure	DS_1	DS_2	DS_3
χ^2	22.2	5.6	0
p-value	2.4E-06	0.02	1
Sep_{norm}	1.39	1.31	1.27

Consistent with Theorem 1, if $\chi^2(DS_1) > \chi^2(DS_2) > \chi^2(DS_3)$, it follows that more distinct clusters can be formed in DS_1 than in DS_2 and DS_3, as demonstrated by the inequality $\text{Sep}_{\text{norm}}(DS_1) > \text{Sep}_{\text{norm}}(DS_2) > \text{Sep}_{\text{norm}}(DS_3)$. Notably, DS_3 represents completely randomized data since each cell count equals its expected value, akin to a single, indivisible cluster from a Gaussian distribution in numerical data, which suggests an absence of clustering structure. Furthermore, our method provides a p-value, inherently establishing a statistically sound threshold (e.g., p-value≤ 0.01), at which our method can determine that only DS_1 is clusterable. In DS_1, 75 students can be perfectly grouped into two clusters: "Good-in-Physics & Good-in-Math" and "Poor-in-Physics & Poor-in-Math". However, in DS_2 and DS_3, with only 65 and 55 students, respectively, able to be grouped, all grades are not considered to have sufficient association.

Table 1 The properties of 18 UCI data sets

Data set	*Abbr.*	#Objects	#Attributes	#Categories
Lenses	Ls	24	4	9
Lung Cancer	Lc	32	56	159
Soybean (Small)	So	47	21	58
Zoo	Zo	101	16	36
Promoter Sequences	Ps	106	57	228
Hayes-Roth	Hr	132	4	15
Lymphography	Ly	148	18	59
Heart Disease	Hd	303	13	57
Solar Flare	Sf	323	9	25
Primary Tumor	Pt	339	17	42
Dermatology	De	366	33	129
House Votes	Hv	435	16	48
Balance Scale	Bs	625	4	20
Credit Approval	Ca	690	9	45
Breast Cancer	Bc	699	9	90
Mammographic Mass	Mm	824	4	18
Tic-Tac-Toe	Tt	958	9	27
Car Evaluation	Ce	1728	6	21

4.2 Data sets and validation strategies

To illustrate and evaluate TestCat, we initiate our study with a selection of commonly used categorical data sets from the UCI repository [42], many of which are likely to have natural clustering structure drawn from various subject areas including life, social, physical, financial fields. The properties of 18 UCI data sets utilized in our study are outlined in Table 1.

Since there are still no recognized benchmark data sets and validation methodology for evaluating the clusterability of categorical data sets, we adopt the following strategy in the performance comparison. Our approach presupposes that all data sets from UCI are clusterable, while their corresponding randomized data sets lack a clustering structure. The details of our validation strategy are elaborated below.

For each original data set (ODS), we generate its corresponding randomized data set (CRDS) (generated as explained in Sect. 3.3). If a data set has a clustering structure, an effective clusterability evaluation method should be able to recognize its ODS as being clusterable while identifying its CRDS as being unclusterable. That is, each ODS is a true positive (clusterable) and all their CRDSs are false positives (unclusterable). Conversely, for a data set devoid of a clustering structure, it is desirable to correctly identify both its ODS and CRDS as being unclusterable. In all scenarios, the ability to discern random data constitutes an essential requirement for any trustworthy clusterability evaluation method.

Since CRDS is not unique, we generate a representative CRDS from a pool of multiple CRDSs for each ODS: Initially, we generate 101 randomized data sets for each ODS. The distribution of p-values (calculated by TestCat) of these 101 randomized data sets is shown in Supplementary Figure 1. In most cases, the p-values of these 101 CRDSs exceed 0.01, as displayed in Supplementary Table 1. From these pool, we obtain a median p-value. Subsequently, we generate additional CRDSs. The chosen CRDS is the first one sequentially generated over multiple runs, with its p-value deviating no more than 0.05 from the median.

4.3 Compared methods

Since there are still no clusterability evaluation methods that are specially developed for categorical data, we adopt and repurpose methodologies typically applicable to numerical data [11]. These methods consist of two key elements: obtaining one-dimensional transformed values and conducting multimodality tests. Dip test (abbreviated as Dip) and Silverman test (abbreviated as Silv) are among the commonly used multimodality tests [11, 13]. To obtain the one-dimensional representation, one method is to acquire pairwise distance through measures specifically crafted for categorical data, and the other is to employ similarity-preserved dimensionality reduction. In our experiment, we trialed 20 different categorical distance measures [21], and subsequently applied Dip and Silv tests, denoted as Dip-dist and Silv-dist, respectively. When deploying similarity-preserved dimensionality reduction techniques, we utilized MDS [37], tSNE [38] and UMAP [39], denoted as MDS Dip/ Silv, tSNE Dip/ Silv, and UMAP Dip/ Silv, respectively.

Here, we clarify the inherent stochastic factors in some of these dimensionality reduction methods and their implications on our experimental setup. For MDS, we leverage a version that incorporates a common heuristic strategy to select the solution with the lowest stress value from several configurations. The initial configuration is set randomly. Analogously, tSNE and UMAP involve randomness in their optimization processes and the initialization of embedding. Therefore, these techniques might converge to different local optimal solutions, resulting in varied outputs across multiple runs. On the contrary, PCA and SPCA are

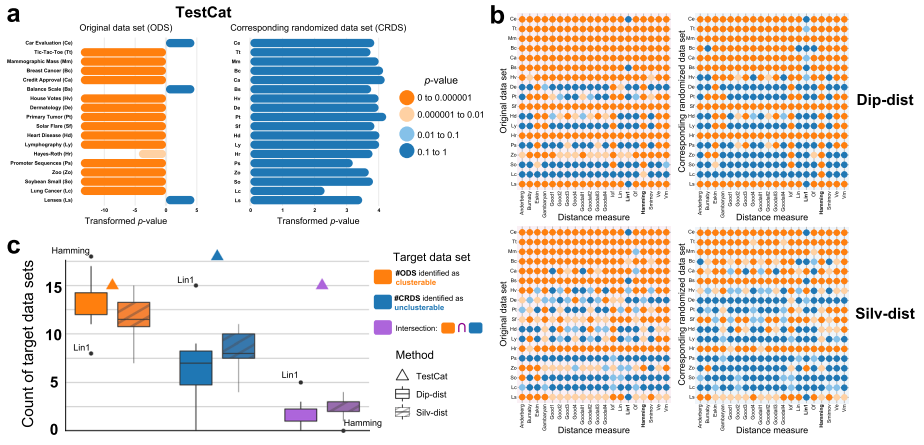

Fig. 3 Identification results of TestCat and existing methods without dimensionality reduction on 18 UCI data sets (including original data sets and corresponding randomized data sets). **a** The barplots of p-values produced by TestCat. To better visualize smaller p-values that approach or equal 0, and to distinguish the significance level of 0.01 through the y-axis ($y = 0$), we use transformed p-values, defined by the formula: $y = \log(p\text{-value}/0.01 + 0.000001) - \log(0.000001)$. **b** Dip and Silverman have been applied to distance values obtained from 20 different categorical distance measures (detailed descriptions of all these measures can be found in reference [43]), hence constituting two sets of comparison methods: Dip-dist and Silv-dist. The outcomes of these comparisons are illustrated in heatmap. **c** We evaluate TestCat against Dip-dist and Silv-dist by counting the number of correctly identified ODSs and CRDSs under the significance level of 0.01. The boxplots describe the count distributions of variants of Dip-dist and Silv-dist derived from the 20 distance measures employed. Outliers are denoted in accordance with the specific distance measure used (color figure online)

deterministic algorithms that invariably generate consistent outputs for a given set of input data. Given these characteristics, in our study, we execute MDS, tSNE, and UMAP multiple times to account for their intrinsic stochasticity. For PCA and SPCA, a single run is sufficient due to their deterministic nature.

In addition to the aforementioned comparative methods, certain advanced embedding methods such as CDE and CDC_DR [29] are designed to transform categorical data into numerical data for clustering tasks. This transformation enables the direct application of conventional clusterability evaluation techniques, which are typically suitable for numerical data. However, these categorical-to-numerical methods tend to generate high-dimensional numerical representations. To tackle this issue in practice, as suggested in reference [13], we employ dimensionality reduction tools such as PCA and sparse principal component analysis (SPCA). These techniques allow us to condense the high-dimensional data into manageable one-dimensional values. We then apply Dip and Silv on that, referred to as PCA Dip/ Silv and SPCA Dip/ Silv, respectively.

4.4 Performance comparison results

We now turn to the evaluation and comparison of TestCat with other clusterability evaluation methods, employing the validation strategy described above. Our investigation reveals that, using TestCat for evaluation, most of these UCI data sets are identified as being clusterable and all their CRDSs are identified as being unclusterable (see Fig. 3a and Table 2), a

Table 2 The clusterability analysis results of TestCat on 18 UCI data sets

Data set	ODS		CRDS	
	#Pairs	p-value	#Pairs	p-value
Ls	0 (0%)	**1**	2 (6.7%)	**0.32**
Lc	1167 (9.4%)	**5E-111**	570 (4.6%)	**0.10**
So	472 (29.8%)	**7E-248**	62 (3.9%)	**0.45**
Zo	306 (51.0%)	**2E-267**	26 (4.3%)	**0.40**
Ps	1537 (6.0%)	**5E-26**	1141 (4.5%)	**0.24**
Hr	21 (25.0%)	**1E-4**	2 (2.4%)	**0.45**
Ly	358 (22.2%)	**8E-195**	82 (5.1%)	**0.55**
Hd	250 (17.0%)	**1E-79**	63 (4.3%)	**0.55**
Sf	133 (49.2%)	**5E-180**	15 (5.6%)	**0.47**
Pt	250 (30.3%)	**1E-101**	56 (6.8%)	**0.68**
De	2719 (33.7%)	**0**	330 (4.1%)	**0.54**
Hv	585 (54.2%)	**0**	55 (5.1%)	**0.54**
Bs	0 (0%)	**1**	13 (8.7%)	**0.43**
Ca	219 (26.7%)	**0**	36 (4.4%)	**0.65**
Bc	1296 (36.0%)	**0**	156 (4.3%)	**0.62**
Mm	44 (36.4%)	**2E-194**	5 (4.1%)	**0.54**
Tt	106 (32.7%)	**4E-106**	18 (5.6%)	**0.42**
Ce	0 (0%)	**1**	10 (5.5%)	**0.47**

"#Pairs" indicates the number of pairs of attribute values exhibiting strong positive or negative association, with the percentage shown in parentheses representing the proportion of such associated pairs. The distribution of these #Pairs from CRDSs (101 runs) for each ODS is provided in Supplementary Figure 1b

conclusion not necessarily guaranteed by existing methods (see Fig. 3c, Fig. 4, Supplementary Figure 2 and Supplementary Figure 3). More details are as follows.

Initially, we examine existing methods that circumvent dimensionality reduction. As depicted in Fig. 3b, none of these methods manage to identify all CRDSs as being unclusterable. Among twenty measures evaluated, only six successfully identify more than half of all target CRDSs (#CRDS $\geq$ 10): both Dip-dist and Silv-dist methods based on Lin1, along with Silv-dist method based on Good1, Good4, Goddall1, Goddall4, and Of. When it comes to identifying more than half of all ODSs as being clusterable (#ODS $\geq$ 10), most methods accomplish this feat, except for Dip-dist using Lin1 and Silv-dist using Good1, Goddall1, and Of. Intriguingly, Dip-dist method utilizing Hamming distance correctly identifies every ODS but remarkably fails to recognize any CRDSs. As demonstrated in Fig. 3c), no single method is able to simultaneously and correctly recognize all ODSs and CRDSs. More precisely, we can employ the count of data sets in the intersection of correctly identified ODS and CRDS as a quantitative indicator. Even the most effective existing method, Dip-dist based on Lin1, could only identify an intersection set with five data sets: Zo, Hv, Ca, Bc, and Tt.

In light of above experimental results, Hamming and Lin1 distances display distinguished performance with respect to at least one or two of the three count metrics (as represented by the outliers in Fig. 3c): number of correctly identified ODS, number of correctly identified CRDS, and the size of intersection set. Nevertheless, all these methods cannot correctly identify all CRDSs, thus failing to meet the trustworthy standard required to discern ran-

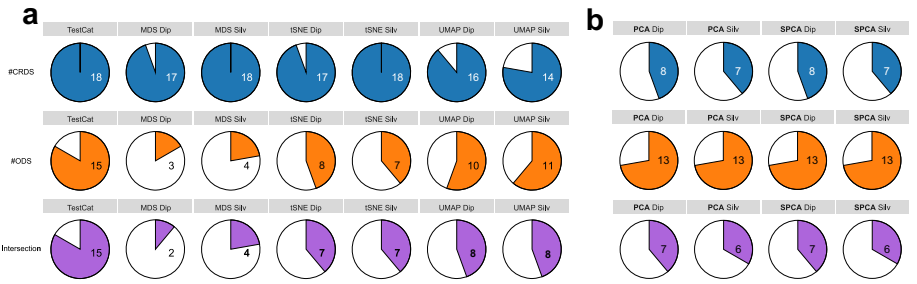

Fig. 4 Count of correctly identified data sets by using TestCat and compared methods. **a** Compared methods via dimensionality reduction (running 101 times for each data set) based on **Hamming** distance. The resulting median p-value from 101 runs is used for determining whether each target data set is clusterable. The experimental results based on Lin1 distance are displayed in Supplementary Figure 2. **b** Categorical-to-Numerical methods via CDC_DR embedding. We implemented the CDC_DR embedding on each categorical data set to generate its numerical representation. Following this transformation, we utilized PCA or SPCA to further condense this numerical data. The experimental results based on CDE embedding are displayed in Supplementary Figure 3

dom data. Consequently, we leverage Hamming distance and Lin1 in the methods based on dimensionality reduction in the performance comparison.

For those methods based on dimensionality reduction, we execute MDS/ tSNE/UMAP 101 times to account for their stochasticity. We then use the median p-value (see Supplementary Table 2.1 and Supplementary Table 2.2) as the identification result of each data set (the detailed p-values of each method on ODSs and CRDSs are displayed in Supplementary Figure 4.1a and Supplementary Figure 4.2a). As demonstrated in Fig. 4a and Supplementary Figure 2, several methods, including MDS Silv based on Hamming distance and Lin1, tSNE Silv based on Hamming distance, and MDS Dip based on Lin1, are capable of identifying all CRDSs correctly. However, these methods do not achieve a satisfactory level of success on identifying most ODSs as being clusterable, thereby revealing their potential limitations as clusterability evaluation methods. Moreover, as displayed in Supplementary Figure 4.1b and Supplementary Figure 4.2b, the methods based on dimensionality reduction generally tend to classify only six data sets, namely So, Zo, De, Hv, Ca, and Bc, as being clusterable.

For the transformation of categorical data into numerical data, it is important to note that such process might result in the loss of some information. As shown in our experiments (see Supplementary Table 3.1 and Supplementary Table 3.2), these embedding-based methods did not achieve better performance than the dimensionality reduction-based methods. As illustrated in Fig. 4b and Supplementary Figure 3, they can identify the majority of ODSs as being clusterable, yet struggle to reliably recognize most CRDSs as being unclusterable. Furthermore, all embedding-based methods also fall short in identifying more than half of the target data sets as indicated by the intersection set.

In order to elucidate rationale of our validation strategy used above, we employ visualization tools to inspect different types of data sets. As depicted in Fig. 5, we present the visualization results on eight data sets by using three tools: scatter plots in reduced two-dimensions space (Fig. 5a), iVAT plots (Fig. 5b), and dissimilarity plots (Fig. 5c). From the visual representations of ODSs and CRDSs, it can be observed that not all ODSs (e.g., Ls, Bs, Ce) has clearly observed clustering structure. This partially illustrate why our method regards the ODS of Ls, Bs and Ce as being unclusterable. The visualization results based on iVAT for the remaining ten data sets are provided in Supplementary Figure 5.

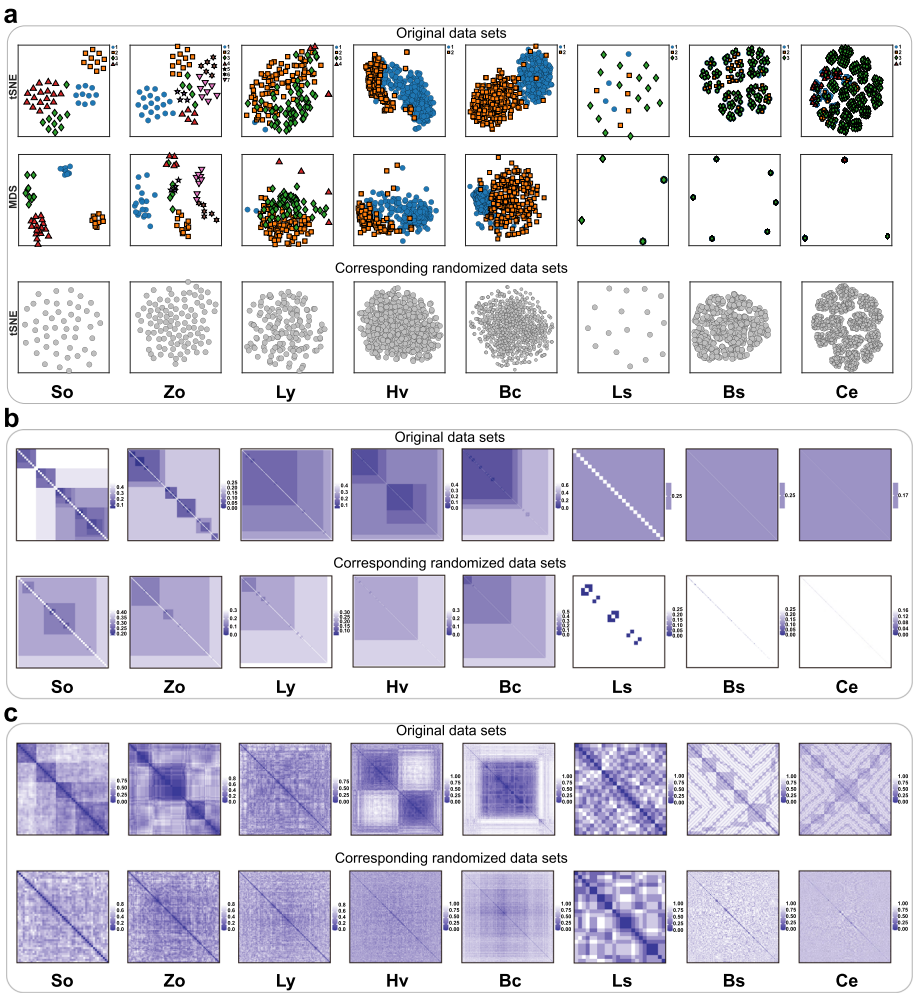

Fig. 5 Illustration of clustering structure underlying 8 UCI categorical data sets by using visual assessment. All plots are derived from the Hamming distances between objects in the data sets. Here, for each original data set, we generate its corresponding randomized data set, which undoubtedly should have no clustering structure (detailed procedures for generating the random data are presented in Sect. 3.3). **a** Scatter plots of tSNE and MDS, where different colors/ shapes represent the class labels provided in the original data sets. Note that duplicate objects have been removed before running tSNE and MDS. Both **b** iVAT plots and **c** Dissimilarity plots display the reordered distances obtained through R Package "seriation". Potential clusters can be identified as multiple densely shaded blocks along the diagonal, where each square is large enough to accommodate a sufficient number of objects. The results from applying iVAT plots to the other 10 UCI categorical data sets are displayed in Supplementary Figure 5 (color figure online)

4.5 Analysis of TestCat

4.5.1 Uniformity

Our clusterability evaluation method, while differing from widely used methods for numerical data in terms of its null hypothesis and test statistic, still relies on the principle of

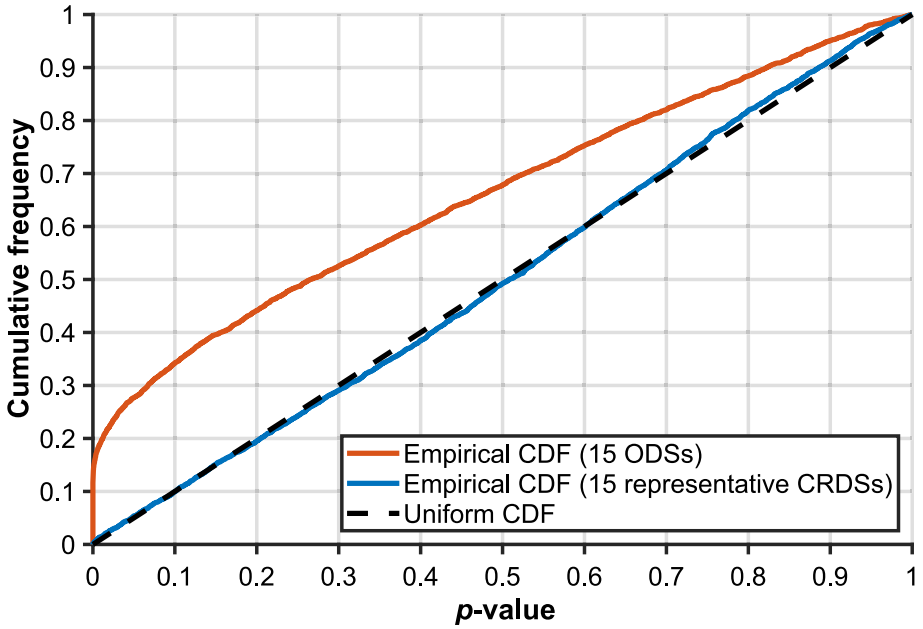

Fig. 6 Comparison of empirical CDFs for p-values of attribute pairs with a Uniform CDF. Such comparisons for each data set are provided in Supplementary Figure 6 (color figure online)

uniformity. For randomized categorical data, the association strength between any attribute pairs is expected to be uniformly distributed across the [0, 1] interval, provided the association strength is arbitrary. This principle holds regardless of the data's scale, ensuring that the p-values for attribute pairs are unbiased and reflect random behavior when there is no conclusive evidence to suggest the presence of a cluster structure.

To validate this, we re-analyzed 15 of the 18 real-world data sets (excluding Ls, Bs, and Ce) that were identified as being clusterable by our TestCat in previous experiments. We compared the empirical cumulative distribution functions (CDFs) of these p-values (prior to aggregation into the final p-value) from the set of ODSs and their representative CRDSs to a theoretical Uniform CDF over the interval [0, 1], as shown in Fig. 6. According to the Kolmogorov–Smirnov test, the CDF from the CRDSs showed no significant difference from the Uniform distribution at a significance level of 0.05 or 0.01. In contrast, the CDF from the ODSs exhibited significant deviations from uniformity. This confirms that the p-values measuring the association strength between attribute pairs from randomized data are approximately uniformly distributed.

Furthermore, our TestCat accurately captures significant associations, aggregating from the attribute-pair level to the data set level, with the overall p-value potentially skewing toward smaller or larger values depending on the observed CDF. Specifically, when strong associations are present among attribute pairs, resulting in many CDF values gathering at the lower tail, as shown in Fig. 6, the TestCat derives an overall p-value that tends to skew toward smaller values. However, such skew is unlikely to occur when the associations among attribute pairs in randomized data are uniformly distributed, ensuring that extreme small final p-values are avoided.

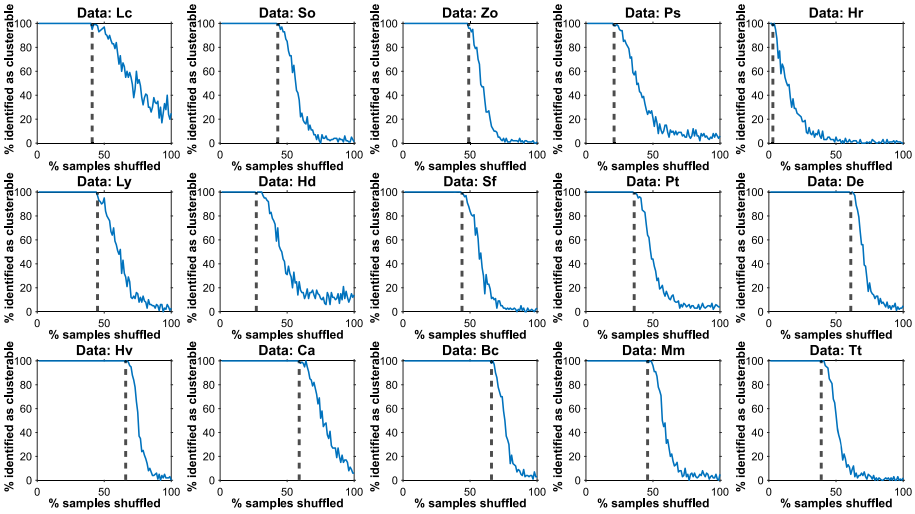

Fig. 7 Robustness of TestCat in determining clusterability against increasing randomness

4.5.2 Robustness

A good clusterability evaluation method should demonstrate robustness across various randomized data sets, correctly identifying them as being unclusterable in most circumstances. Our approach exemplifies this property using the association strength among attributes. As illustrated in Supplementary Figure 1b, the majority of strongly associated attribute pairs fall within the relatively narrow range across the 101 CRDSs for each ODS. In contrast, the proportion of associated attribute pairs for clusterable ODS markedly exceed this value, displaying a broader distribution range. As a result, the p-value for randomized data sets consistently exceed that of the clusterable ODS, implying a clear distinction between the ODS and these randomized data sets (refer to Fig. 3a).

As depicted in Supplementary Figure 7a, using a representative CRDS for each ODS, we observe a clear distinction in the number of positively associated attribute value pairs between the ODS and CRDS across eight data sets. These situations can be classified into two types: one where the number of associated pairs of the ODS markedly exceeds that of the CRDS and another where the number of associated pairs of the ODS is zero (data sets Ls, Bs and Ce). The former scenario aligns with our criteria for identifying clusterable ODS, while the latter leads our method to compute a p-value of 1, marking them as unclusterable (an outcome that is consistent with the evidence provided by iVAT). Such plots on the number of associated pairs for the remaining data sets are displayed in Supplementary Figure 8. In all these cases, the number of associated pairs for the ODS is evidently greater than that for the CRDS, even though the difference is less pronounced in the data sets Lc and Ps. More specifically, as shown in Supplementary Figure 7b, the proportion of associated attribute pairs for clusterable ODSs exceeds 20% in most cases. In contrast, the proportion for CRDSs is below 8.7% in all cases.

In addition to demonstrating TestCat's robustness in distinguishing between the clusterability of ODS and CRDS, we further extend our analysis to a simulation study examining TestCat's performance under conditions of partial randomness. We generate locally randomized data in the same manner as described in Sect. 3.3, but instead of shuffling all objects

Fig. 8 Runtime for 50 iterations of each method across 18 UCI data sets. The values within each cell of the heatmap represent the actual runtime in seconds. The coloring of the heatmap is determined by the normalized runtime: for each data set, the runtime is scaled to the range (0,1] by dividing it by the maximum runtime. The corresponding color scale is illustrated in the legend on the right side (color figure online)

across attributes, we shuffle only a fraction of them. As shown on the horizontal axis of Fig. 7, we progressively shuffle 1%, 2%, and up to 100% of the objects, with 100% corresponding to the generation of a CRDS. Each shuffle is independently repeated 100 times, and the vertical axis represents the proportion of cases that TestCat identifies as being clusterable. The results indicate that when only a small fraction of objects are randomized, the derived p-values remain below the significance level of 0.01, identifying the cases as being clusterable with high accuracy. As more objects are shuffled, the strength of associations diminishes, and TestCat gradually fails to accurately detect clustering structures, classifying more cases as being unclusterable. When 100% of the objects are shuffled (equivalent to generating CRDS), these cases are consistently identified as being unclusterable. Notably, even when nearly 50% or more of the objects are shuffled, TestCat can still perfectly identify the cases as being clusterable with 100% accuracy (as shown by the horizontal line in each plot of Fig. 7) based on most of the clusterable data sets: Ls, So, Zo, Ly, Sf, De, Hv, Ca, Bc, Mm and Tt. This highlights TestCat's robustness against increasing randomness across a wide range of different clusterable data sets.

4.5.3 Efficiency

Efficiency is crucial when a clusterability evaluation method is applied to large-scale data sets. To underscore this point, we assessed TestCat's runtime on 18 UCI data sets and compared it with other methods, specifically those based on dimensionality reduction. As illustrated in

Fig. 8, TestCat showcased superior speed, topping the charts on eight specific data sets: Ls, Hr, Ly, Sf, Bs, Ca, Mm, and Ce. Notably, on 15 data sets, excluding Lc, Ps, and De, TestCat executed in less than 25% of the duration required by the slowest competitor. Overall, in comparison with the fastest benchmark, MDS Dip, TestCat achieved a runtime that was approximately 40% of its execution time.

4.6 Code and data availability

TestCat is available on GitHub (https://github.com/hulianyu/TestCat). Detailed guidelines for its implementation, as well as for the execution of the compared methods, are provided. Information on the parameters used in the codes of the compared methods is presented in Supplementary Note 1.

The complete collection of 18 UCI data sets employed in our analysis is publicly accessible at https://archive.ics.uci.edu/datasets (Attribute Type = Categorical), with their properties illustrated in Table 1. All utilized CRDSs, in addition to the numerical representations for each ODS and CRDS, can be found in the same GitHub repository hosting TestCat.

5 Conclusions and Discussions

Clusterability evaluation plays a crucial role in cluster analysis. The importance of this step is underscored by its ability to authenticate the validity of clustering results, ensuring that identified patterns truly originate from the data rather than being mere by-product of the clustering process. More explicitly, if data are determined as being unclusterable, the use of any clustering algorithm would invariably lead to meaningless results.

To address the clusterability evaluation issue for categorical data, we introduce a testing-based method by employing the association relationship among categorical attributes. To the best of our knowledge, this is the first piece of work to tackle the clusterability evaluation problem for categorical data. Compared to existing solutions that are not specially developed for categorical data, our TestCat approach not only demonstrates superior performance but also solves such a challenging issue in an intuitive, simple and elegant manner.

It is undeniable that our method harbors certain limitations: (1) The independence assumption [44] among attribute pairs, as well as the use of the Chi-squared approximation (further discussed in Supplementary Note 2), generally do not hold. If we do not adhere to these assumptions, the deviation of an analytical p-value for significance assessment would be a challenging issue. (2) The Chi-squared test may not be the best choice for association testing in the context of clusterability evaluation, we need to further check other more appropriate methods. (3) Association testing may lack validity for data structures with intricate attribute interactions. For instance, in the Bs (Balance Scale) data set, cluster formation may rely on multi-attribute interactions beyond just pairwise associations. Additionally, in the Ce (Car Evaluation) data set, the absence of associations between individual attributes does not necessarily imply there are no clusters, as these hierarchical attributes could be strongly associated with a cluster variable. These scenarios, as detailed in Supplementary Note 3, indicate the need for developing alternative methods capable of capturing such complex relationships.

Finally, our research efforts in this article hint at several aspects. First, it highlights the importance of the clusterability evaluation issue for categorical data, which is overlooked during the past decades. Second, we empirically demonstrate the fact that existing solutions for numeric data cannot solve this problem very well, indicating that it is still necessary to

develop new algorithms toward this direction. Finally, there are many directions for future research, ranging from the refinement of TestCat to the development of brand-new methods.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10115-024-02317-x.

Acknowledgements This work has been partially supported by the Natural Science Foundation of China under Grant No. 62472064.

References

1. Kleinberg J (2002) An impossibility theorem for clustering. Adv Neural Inf Process Syst 15
2. Pelillo M (2009) What is a cluster? perspectives from game theory. In: Proceedings of the NIPS Workshop on Clustering Theory
3. Hennig C (2015) What are the true clusters? Pattern Recogn Lett 64:53–62
4. Thomann P, Steinwart I, Schmid N (2015) Towards an axiomatic approach to hierarchical clustering of measures. J Mach Learn Res 16(59):1949–2002
5. Ben-David S, Ackerman M (2008) Measures of clustering quality: a working set of axioms for clustering. Adv Neural Inf Process Syst 21
6. Garza-Fabre M, Handl J, José-García A (2023) Evolutionary multiobjective clustering over multiple conflicting data views. IEEE Trans Evol Comput 27(4):817–831
7. Von Luxburg U, Williamson RC, Guyon I (2012) Clustering: science or art? In: Proceedings of ICML Workshop on Unsupervised and Transfer Learning, 65–79
8. Arbelaitz O, Gurrutxaga I, Muguerza J, Pérez JM, Perona I (2013) An extensive comparative study of cluster validity indices. Pattern Recogn 46(1):243–256
9. Diallo AF, Patras P (2024) Deciphering clusters with a deterministic measure of clustering tendency. IEEE Trans Knowl Data Eng 36(4):1489–1501
10. Ackerman M, Ben-David S (2009) Clusterability: a theoretical study. Int Conf Artif Intell Stat 5:1–8
11. Adolfsson A, Ackerman M, Brownstein NC (2019) To cluster, or not to cluster: an analysis of clusterability methods. Pattern Recogn 88:13–26
12. Ahmadi S, Awasthi P, Khuller S, Kleindessner M, Morgenstern J, Sukprasert P, Vakilian A (2022) Individual preference stability for clustering. In: International Conference on Machine Learning, 197–246
13. Laborde J, Stewart PA, Chen Z, Chen YA, Brownstein NC (2023) Sparse clusterability: testing for cluster structure in high dimensions. BMC Bioinformatics 24(1):1–27
14. Cheng M-Y, Hall P (1998) Calibrating the excess mass and dip tests of modality. J R Stat Soc: Ser B (Statistical Methodology) 60(3):579–589
15. Silverman BW (1981) Using kernel density estimates to investigate multimodality. J Roy Stat Soc: Ser B (Methodol) 43(1):97–99
16. Havens TC, Bezdek JC (2011) An efficient formulation of the improved visual assessment of cluster tendency (IVAT) algorithm. IEEE Trans Knowl Data Eng 24(5):813–822
17. Hahsler M, Hornik K (2011) Dissimilarity plots: a visual exploration tool for partitional clustering. J Comput Graph Stat 20(2):335–354
18. Agresti A (2012) Categorical Data Analysis, vol 792. John Wiley & Sons, Hoboken
19. Couper MP (2017) New developments in survey data collection. Ann Rev Sociol 43:121–145
20. Vasaikar SV, Savage AK, Gong Q, Swanson E, Talla A, Lord C, Heubeck AT, Reading J, Graybuck LT, Meijer P (2023) A comprehensive platform for analyzing longitudinal multi-omics data. Nat Commun 14(1):1684
21. Boriah S, Chandola V, Kumar V (2008) Similarity measures for categorical data: a comparative evaluation. In: SIAM International Conference on Data Mining, 243–254
22. Naouali S, Ben Salem S, Chtourou Z (2020) Clustering categorical data: a survey. Int J Inf Technol Decis Mak 19(01):49–96
23. Liu H, Chen J, Dy J, Fu Y (2023) Transforming complex problems into k-means solutions. IEEE Trans Pattern Anal Mach Intell 45(7):9149–9168
24. Bouguessa M (2015) Clustering categorical data in projected spaces. Data Min Knowl Disc 29:3–38
25. Cesario E, Manco G, Ortale R (2007) Top-down parameter-free clustering of high-dimensional categorical data. IEEE Trans Knowl Data Eng 19(12):1607–1624
26. Zhang Y, Cheung Y-M (2022) Learnable weighting of intra-attribute distances for categorical data clustering with nominal and ordinal attributes. IEEE Trans Pattern Anal Mach Intell 44(7):3560–3576

27. Xiao Y, Huang C, Huang J, Kaku I, Xu Y (2019) Optimal mathematical programming and variable neighborhood search for k-modes categorical data clustering. Pattern Recogn 90:183–195
28. Jian S, Pang G, Cao L, Lu K, Gao H (2019) Cure: flexible categorical data representation by hierarchical coupling learning. IEEE Trans Knowl Data Eng 31(5):853–866
29. Bai L, Liang J (2022) A categorical data clustering framework on graph representation. Pattern Recogn 128:108694
30. Barbará D, Li Y, Couto J (2002) Coolcat: an entropy-based algorithm for categorical clustering. In: Proceedings of the Eleventh International Conference on Information and Knowledge Management, pp. 582–589
31. Li T, Ma S, Ogihara M (2004) Entropy-based criterion in categorical clustering. In: Proceedings of the Twenty-first International Conference on Machine Learning, p. 68
32. Fisher DH (1987) Knowledge acquisition via incremental conceptual clustering. Mach Learn 2(2):139–172
33. Mirkin B (2001) Reinterpreting the category utility function. Mach Learn 45(2):219–228
34. Epter S, Krishnamoorthy M, Zaki M (1999) Clusterability detection and initial seed selection in large datasets. In: The International Conference on Knowledge Discovery in Databases, vol. 7
35. Liu Y, Hayes DN, Nobel A, Marron JS (2008) Statistical significance of clustering for high-dimension, low-sample size data. J Am Stat Assoc 103(483):1281–1293
36. Dubes RC, Zeng G (1987) A test for spatial homogeneity in cluster analysis. J Classif 4:33–56
37. Leeuw J, Mair P (2009) Multidimensional scaling using majorization: Smacof in r. J Stat Softw 31(3):1–30
38. Maaten L, Hinton G (2008) Visualizing data using T-SNE. J Mach Learn Res 9(11)
39. Becht E, McInnes L, Healy J, Dutertre C-A, Kwok IW, Ng LG, Ginhoux F, Newell EW (2019) Dimensionality reduction for visualizing single-cell data using UMAP. Nat Biotechnol 37(1):38–44
40. Hayes-Roth B, Hayes-Roth F (1977) Concept learning and the recognition and classification of exemplars. J Verbal Learn Verbal Behav 16(3):321–338
41. Cinar O, Viechtbauer W (2022) The poolr package for combining independent and dependent p values. J Stat Softw 101:1–42
42. Dua D, Graff C (2019) UCI Machine Learning Repository
43. Sulc Z, Cibulkova J, Rezankova H (2022) Nomclust 2.0: an R package for hierarchical clustering of objects characterized by nominal variables. Comput Statistics 37(5):2161–2184
44. Ferrari A (2019) A note on sum and difference of correlated chi-squared variables. arXiv preprint arXiv:1906.09982

Lianyu Hu received the MS degree in computer science from Ningbo University, China, in 2019. He is currently working toward the PhD degree in the School of Software at Dalian University of Technology. His current research interests include machine learning and data mining.

Junjie Dong received the BS degrees in chemistry from Dalian University of Technology and University of Leicester in 2022, respectively. He is currently working toward the MS degree in the School of Software at Dalian University of Technology. His research interests include bioinformatics and data mining.

Mudi Jiang received the MS degree in software engineering from Dalian University of Technology, China, in 2023. He is currently working toward the PhD degree in the School of Software at the same university. His current research interests include data mining and its applications.

Yan Liu received the MS degree in applied statistics from Dongbei University of Finance and Economics in 2018, and the PhD degree from Dalian University of Technology in 2022. She is currently a lecturer in the School of Software Engineering, Dalian University. Her research interests include bioinformatics and data mining.

Zengyou He received the BS, MS, and PhD degrees in computer science from Harbin Institute of Technology, China, in 2000, 2002, and 2006, respectively. He was a research associate in the Department of Electronic and Computer Engineering, Hong Kong University of Science and Technology, from February 2007 to February 2010. He is currently a professor in the School of Software, Dalian University of Technology. His research interest include data mining and bioinformatics.