entropy

MDPI

Article

Nonparametric Clustering of Mixed Data Using Modified Chi-Squared Tests

Yawen Xu *, Xin Gao and Xiaogang Wang

Department of Mathematics and Statistics, York University, Toronto, ON M3J 1P3, Canada
* Correspondence: yawenxu@yorku.ca

Abstract: We propose a non-parametric method to cluster mixed data containing both continuous and discrete random variables. The product space of the continuous and discrete sample space is transformed into a new product space based on adaptive quantization on the continuous part. Detection of cluster patterns on the product space is determined locally by using a weighted modified chi-squared test. Our algorithm does not require any user input since the number of clusters is determined automatically by data. Simulation studies and real data analysis results show that our proposed method outperforms the benchmark method, AutoClass, in various settings.

Keywords: clustering; mixed data; non-parametric; weighted chi-squared test

check for updates

Citation: Xu, Y.; Gao, X.; Wang, X. Nonparametric Clustering of Mixed Data Using Modified Chi-Squared Tests. *Entropy* **2022**, *24*, 1749. https://doi.org/10.3390/e24121749

Academic Editor: Friedhelm Schwenker

Received: 4 October 2022
Accepted: 28 November 2022
Published: 29 November 2022

1. Introduction

Mixed data that contain both continuous and discrete data are abundant in scientific research, especially in medical or biological studies. An effective clustering method for mixed data should partition a large complex data set into homogeneous subgroups that are manageable in statistical inference. Clustering methods thus have a wide range of applications in almost all scientific studies including financial risk analysis, genetic analysis, and medical studies. They are essential tools in analyzing large data sets.

Most of the clustering methods in the literature have been mainly focused on either continuous data or categorical data alone. The K-means algorithm has been widely used in industrial applications for a long time. Detailed descriptions and discussions can be found in Kaufman and Rousseeuw (2009) [1]. Non-Euclidean distances such as the Manhattan distance or Mahoblis distance have also been used. Model-based clustering methods for continuous data have been proposed in the literature, see for example Banfield and Raftery (1993) [2]. One of the most prominent methods in parametric clustering based on a mixture model is proposed by Bradley et al. (1998) [3]. The number of clusters and outliers can be handled simultaneously by the mixture model. Fraley and Raftery (1998) [4] propose choosing the number of clusters automatically using the model-based clustering method. For clustering categorical data, there are far fewer reliable methods. K-modes algorithm has been proposed by Huang (1998) [5] to extend the K-means to clustering categorical data. The AutoClass method proposed by Cheeseman and Stutz (1996) [6] is a well-known method in clustering. The AutoClass takes a data set containing both real and discrete-valued attributes and automatically computes the number of clusters and group memberships. This method has been used by NASA and helped to find infra-red stars in the IRAS Low-Resolution Spectral catalog and discovery classes of proteins (Cheeseman and Stutz 1996 [6]).

In clustering mixed data, the main difficulty lies in the fact that continuous and categorical sample spaces are intrinsically different. Although both can be made into metric spaces, the continuous sample space resides on a differentiable manifold while the categorical one is defined entirely on a lattice. Attempts have been made in the literature to combine the two spaces by using a global and general distance function

(Ahmad and Dey 2007 [7]). This naive approach ignores the fact that the two sample spaces are topologically incompatible. Another approach is to apply different clustering algorithms to the continuous and categorical portions separately and combine the results. This approach, however, would sever the intrinsic connection between the continuous and categorical parts of one record. Each record is often assigned to different clusters for the continuous and categorical parts. It is often hard to reconcile this except by expanding the total number of clustering. Not only does this produce a larger than necessary number of clusters for the entire data sets, but a true cluster is also often found being split across many small clusters and renders the results to be very inaccurate. Alternatively, AutoClass combines information across probability spaces. However, the effectiveness of AutoClass depends on the validity of the assumed parametric model. Zhang et al. (2006) [8] showed that both K-modes and AutoClass do not perform very well when applied to benchmark categorical data sets from the UCI machine learning depository. Therefore, there is a need for a non-parametric clustering method for mixed data.

We extend the work by Zhang et al. (2006) [8] to cluster mixed data by using adaptive quantization of the continuous sample space. The quantization process was developed in the 1950s and it partitions the sample space through a discrete-valued map (Gersho and Gray 1992) [9]. For univariate cases, quantization is known as vector quantization and it is the fundamental process for converting analog signals or information into digital forms (Gersho and Gray, 1992) [9]. It has been used in studying pricing in finance as well as engineering. Theoretical properties of quantization in probability distributions can be found in Graf and Luschgy (2000) [10]. The process of clustering mixed data is then performed on the quantized product space. The key idea is inspired by the fact that any manifold can be locally modeled by a Euclidian space. Therefore, each neighborhood in the transformed product space can be locally characterized as a fine grid endowed with a Hamming Distance. The Hamming Distance is widely used in information and coding theory (Roman 1992 [11]; Laboulias et al., 2002 [12]). The statistical significance of a detected cluster is determined by a weighted local Chi-squared test. The advantage of our proposed method over AutoClass is demonstrated in simulations and by using two benchmark data sets from the UCI machine learning depository.

This paper is organized as follows. The method is proposed in Section 2. The clustering algorithm is presented in Section 3. Simulation results are provided in Section 4.

2. Clustering Methodology

In this section, we introduce quantization of the mixed sample space on which we adopt the Hamming Distance function to measure the relative positions of two data points. We also define a distance vector and an optimal separation point which are essential to measuring spatial patterns as well as the size of any detected clusters. Separation points are introduced to extract detected cluster patterns.

2.1. Joint Sample Space of Mixed Data

Consider a general data structure for a mixed data set with p nominal categorical attributes and q continuous attributes. The categorical sample space is defined on $\Omega_p = R^p$ while the continuous one is defined on Ω_q. The product space for mixed data is then defined on the product space $\Omega_p \otimes \Omega_q$. The sample size is denoted by n.

The categorical part of mixed data is represented by $\mathbf{X} = (X_i^j)$, with $i = 1, 2, \ldots, n$ and $j = 1, \cdots, p$. Furthermore, row and column vectors in the categorical portion are denoted by $\mathbf{X}_i^{[\cdot]}$ and $\mathbf{X}_{[\cdot]}^j$. The j^{th} categorical attribute is categorized by m_j levels defined by set $A_j = (a_{j1}, \cdots, a_{jm_j}), j = 1, \cdots, p$.

We denote the continuous part of a mixed sample with size n by $\mathbf{Z} = (Z_i^k)$, with $i = 1, 2, \ldots, n$ and $k = 1, \cdots, q$. Furthermore, we denote the row and column vectors in the categorical portion by $\mathbf{Z}_i^{[\cdot]}$ and $\mathbf{Z}_{[\cdot]}^k$. The k^{th} attribute is a continuous random variable.

2.2. Quantization of Continuous Sample Space

Continuous data and discrete data are fundamentally different. Although the description provided by the continuous portion can be very detailed, it could carry excessive information that is not important for the clustering purpose. Furthermore, any pattern derived from the categorical part is based on a much coarse topology than the continuous counterpart. Since it is impossible to define a meaningful and objective manifold from a coarse data structure, the continuous one then must be mapped into a grid that is compatible with the relatively coarse topology from the categorical one.

The quantization is achieved in two steps. Firstly for observed realization z_i^j, continuous data are mapped onto the unit interval between 0 to 1 by applying the following formula:

$$\tilde{z}_i^k = \frac{z_i^k - z_{min}^k}{z_{max}^k - z_{min}^k}, \quad k = 1, ..., q; i = 1, ..., n,$$

where z_{min}^k and z_{max}^k represent the minimum and maximum values of the k column. Secondly, for the standardized observations, the continuous random variable is then mapped or quantized into a discrete random variable with M levels in the following way:

$$Q(\tilde{z}_i^k) = m, \quad if \quad (m-1)/M \leq \tilde{z}_i^k < m/M,$$

where $m = 1, 2, \cdots, M$, where M can be any positive integer value. Different numerical values of M could have an impact on the quality of quantization and consequently the clustering result. A finer quantization grid might not be useful and could be more computationally intensive than a coarse one.

The number of levels M can be difficult to specify by a user with no prior information. Thus, we propose to choose level M adaptively by using F statistics based on the clustering results.

For any fixed numerical value of M, we perform an ANOVA test by treating each cluster as a separate group by using the generated clustering memberships. The F-statistic associated with the ANOVA test is recorded for different values of M. We then set the optimal numerical value for M by selecting the value that corresponds to the largest F-statistic computed before. Numerical results of the quantization level will be illustrated in Section 4.1.

2.3. Distance Vectors on Quantized Product Space

We use Hamming Distance (HD) to measure the relative separation of two categorical data points. To be more specific, for any two positions in the categorical sample space Ω_p, $\mathbf{Q}_h^{[\cdot]} = (Q_h^{[1]}, \cdots, Q_h^{[p]})$ and $\mathbf{Q}_i^{[\cdot]} = (Q_i^{[1]}, \cdots, Q_i^{[p]})$, the HD between $Q_h^{[j]}$ and $Q_i^{[j]}$ on the jth attribute is

$$d(Q_h^j, Q_i^j) = \begin{cases} 0 & if \quad Q_h^j = Q_i^j, \\ 1 & if \quad Q_h^j \neq Q_i^j. \end{cases}$$

Further, we define the distance between the two positions, that is, the summation of the distance from each pair of the components. Therefore, we have the following:

$$HD(\mathbf{Q}_h^{[\cdot]}, \mathbf{Q}_i^{[\cdot]}) = \sum_{j=1}^{p} d(\mathbf{Q}_h^j, \mathbf{Q}_i^j).$$

After quantization, the new product space now resides on a high-dimensional grid. Since for a grid, there is no natural origin. We can define a reference point $(\mathbf{S}, \mathbf{T})$ in the quantized product space with $\mathbf{S} = (s_1, \cdots, s_p) \in R^p$ and $\mathbf{T} = (t_1, \cdots, t_q) \in R^q$. For the categorical portion, $HD_C(\mathbf{X}_i, \mathbf{S})$ can take values ranging from 0 to p; and for quantized continuous data, we have $HD_Q(\mathbf{Z}_i, \mathbf{T})$ can take values ranging from 0 to q.

We then define the Distance Vector (DV) based on Hamming distance for the categorical and quantized continuous portion, respectively. We define two individual vectors to record the frequencies of each categorical and quantized continuous distance value accordingly, that is, a $(p+1)$-element vector $DV_C(\mathbf{S})$ for the categorical data and a $(q+1)$-element vector $DV_Q(\mathbf{T})$ for the quantized part. To be more specific, DV_C is defined as

$$DV_C(\mathbf{S}) = (DV_C^{[0]}(\mathbf{S}), DV_C^{[1]}(\mathbf{S}), \cdots, DV_C^{[p]}(\mathbf{S})),$$

and DV_Q is defined as

$$DV_Q(\mathbf{T}) = (DV_Q^{[0]}(\mathbf{T}), DV_Q^{[1]}(\mathbf{T}), \cdots, DV_Q^{[q]}(\mathbf{T})).$$

The j^{th} component in DV_C and h^{th} component in DV_Q are given as the following:

$$DV_C^{[j]}(S) = \sum_{i=1}^{n} \boldsymbol{I}\left[HD_C(\mathbf{X}_i^{[\cdot]}, \mathbf{S}) = j\right], \quad j = 0, 1, \cdots, p;$$

$$DV_Q^{[h]}(T) = \sum_{i=1}^{n} \boldsymbol{I}\left[HD_Q(\mathbf{Q}_i^{[\cdot]}, \mathbf{T}) = h\right], \quad h = 0, 1, \cdots, q;$$

where $\boldsymbol{I}(A)$ is an indicator function that takes value 1 when event A happens and 0 otherwise.

If there is no cluster pattern at all, we would expect a uniform distribution of all possible cases. Then it is equally likely for a randomly chosen data point to take any possible position in the joint sample space. The DV vectors under uniform distribution are referred to as a *uniform* distance vector (UDV). Thus, a UDV records the expected frequencies under the null hypothesis that there are no clustering patterns in the data. Let $\mathbf{X}$ be a categorical portion of data and $\mathbf{Z}$ be a continuous portion of the data from a sample of size n, with each observation having an equal probability of locating at any position on space $\Omega_p \otimes \Omega_q$. The expected value of DV and DV associated with the null hypothesis is denoted by UDV_C, $\mathbf{U} = (U_0, \cdots, U_p)$ for categorical data and UDV_Q, $\mathbf{V} = (V_0, \cdots, V_q)$ for continuous data, respectively.

Zhang et al. (2006) [8] provide the exact form of $UDV_C = \frac{n}{M_1}\mathbf{U}^*$, where $M_1 = \prod_{j=1}^{p} m_j, j = 1, 2, \cdots, p$; m_j is the number of states in set A_j for the j^{th} attribute; and $\mathbf{U}^* = (U_0^*, U_1^*, \cdots, U_p^*)$ with

$$\begin{aligned}
&U_0^* = 1;\\
&U_1^* = (m_1 - 1) + (m_2 - 1) + \cdots + (m_p - 1);\\
&U_2^* = \textstyle\sum_{i<j}^{p}(m_i - 1)(m_j - 1);\\
&\vdots\\
&U_p^* = (m_1 - 1)(m_2 - 1)\cdots(m_p - 1).
\end{aligned}$$

Similarly, we obtain the exact form of the UDV_Q for the quantized continuous part of the data. $UDV_Q = \frac{n}{M_2}\mathbf{V}^*$, where $M_2 = \prod_{j=1}^{q} l_j, j = 1, 2, \cdots, q$; l_j is the the number of levels of quantization for the j^{th} continuous attribute; and $\mathbf{V}^* = (V_0^*, V_1^*, \cdots, V_q^*)$ with

$$\begin{aligned}
&V_0^* = 1;\\
&V_1^* = (l_1 - 1) + (l_2 - 1) + \cdots + (l_q - 1);\\
&V_2^* = \textstyle\sum_{i<j}^{q}(l_i - 1)(l_j - 1);\\
&\vdots\\
&V_q^* = (l_1 - 1)(l_2 - 1)\cdots(l_p - 1).
\end{aligned}$$

2.4. Optimal Separation Point

If the initial starting point is chosen to be the center of one particular cluster, then the frequency of HD should demonstrate a decreasing pattern in a local region as the HD function records the frequency of data points from the center of the cluster and outwards. Small local bumps at the beginning part of the HD curve are expected if the initial starting point deviates slightly from the cluster center. The recorded frequencies might increase afterward when the function begins to record distances from another cluster. Therefore, the valley area indicates a natural place to separate one cluster from the rest. Separation points are, therefore, defined for this identification purpose.

Assume that the categorical data $\mathbf{X}$ and quantized continuous data $\mathbf{Z}$ are not uniformly distributed in the sample space $\Omega_p \otimes \Omega_q$. Let $DV_C(\mathbf{S}) = (DV_C^{[0]}(\mathbf{S}), DV_C^{[1]}(\mathbf{S}), \cdots, DV_C^{[p]}(\mathbf{S}))^T$, $\mathbf{S} \in \Omega_p$ be the collection of all $(p+1)$-element DV_C in the space Ω_p and $DV_Q(\mathbf{T}) = (DV_Q^{[0]}(\mathbf{T}), DV_Q^{[1]}(\mathbf{T}), \cdots, DV_Q^{[q]}(\mathbf{T}))^T$, $\mathbf{T} \in \Omega_q$ be the collection of all $(q+1)$-element DV_Q in the space Ω_q, and let $\mathbf{U} = (U_0, U_1, \cdots, U_p)^T$ be the DV_C vector and $\mathbf{V} = (V_0, V_1, \cdots, V_q)^T$ be the DV_Q vector defined in the previous subsection. For a given distance value $j_C, j_C = 0, 1, \cdots, p$, for categorical distance values and $j_Q, j_Q = 0, 1, \cdots, q$, for quantized continuous distance values, there always exists at least one position $(\mathbf{S},\mathbf{T}) \in \Omega_p \otimes \Omega_q$, such that the frequency at this distance value is larger than the corresponding component, U_j of the UDV_C vector and V_j of the UDV_Q vector.

In order to proceed to a comparison between DV_C and UDV_C and between DV_Q and UDV_C, we introduce a selection criterion for an optimal cut-off r^*. The categorical cut-off point was defined and proved by Zhang et al. (2006) [8]. Because our quantized continuous data behaves as categorical data, we extend that concept to a quantized portion of the data. If the cluster structure is present, the early segment of a DV_C and DV_Q with respect to a data center should contain substantially larger frequencies than the corresponding frequencies of the UDV_C vector and UDV_Q vector. Therefore, the range corresponding frequencies of the UDV_V vector and UDV_Q a vector that is consistently larger than the UDV_C vector and UDV_Q vector gives a reasonable indication of the r. This leads to an optimal r_C^* for the categorical portion of data:

$$r_C^*(\mathbf{S}) = \min_{j_C>0}\{j_C | \frac{DV_C^{[j_C]}(\mathbf{S})}{U_{j_C}} < 1\} - 1, \mathbf{S} \in \Omega_p.$$

Similarly, optimal r_Q^* for the quantized portion of data be:

$$r_Q^*(\mathbf{T}) = \min_{j_Q>1}\{j_Q | \frac{DV_Q^{[j_Q]}(\mathbf{T})}{V_{j_Q}} < 1\} - 1, \mathbf{T} \in \Omega_q.$$

The two quantities are used to identify relatively dense regions in the space of mixed data to help us to extract clusters accurately. Zhang et al. (2006) [8] gave a detailed explanation of the tuition of radius which is the maximum distance of the data points in this cluster to its center.

3. Algorithm

There are two key parts of the algorithm. Firstly, we detect whether there exist any statistically significant clustering patterns. We propose a weighted local Chi-squared test to determine if the observed distance vectors differ significantly from the uniform distance vectors associated with no cluster pattern. Secondly, if the patterns are significant, we further extract the clusters based on the optimal separation strategies described in the previous section.

We consider the null hypothesis H_0: There is no clustering pattern in the data set. The weighted local Chi-squared test statistic $\chi^{2*}_w(\mathbf{S,T})$ is defined as:

$$\begin{aligned}\chi^{2*}_w(\mathbf{S},\mathbf{T}) &= (\tfrac{1}{p}+\tfrac{1}{q})[\tfrac{1}{p}\chi^{2*}_C(\mathbf{S})+\tfrac{1}{q}\chi^{2*}_Q(\mathbf{T})]\\ &= \tfrac{pq}{p+q}\tfrac{1}{p}\chi^{2*}_C(\mathbf{S})+\tfrac{pq}{p+q}\tfrac{1}{q}\chi^{2*}_Q(\mathbf{T})\\ &= \tfrac{q}{p+q}\chi^{2*}_C(\mathbf{S})+\tfrac{p}{p+q}\chi^{2*}_Q(\mathbf{T}), \qquad (\mathbf{S},\mathbf{T})\in\Omega_{p\otimes q}\end{aligned}$$

The weighted local Chi-squared statistic $\chi^{2*}_w(\mathbf{S,T})$ is constructed to address the issue of an unequal number of variables for the continuous and categorical parts. We expect that a large number of variables tend to produce a large numerical value for the modified χ^{2*}_C and χ^{2*}_Q. Therefore, each modified Chi-squared statistic is normalized by its corresponding number of variables for the categorical and continuous parts respectively. To ensure the total of the two weights to equal to 1, we further divide the sum of two normalized modified Chi-squares by the total of the two weights which equals $1/p+1/q$.

Where the categorical part $\chi^{2*}_C(\mathbf{S})$ takes form as:

$$\chi^{2*}_C(\mathbf{S})=\sum_{j=0}^{r^*_C}\frac{(DV^{[j]}_C(\mathbf{S})-U_j)^2}{U_j}+\frac{(\sum_{j=0}^{r^*_C}DV^{[j]}_C(\mathbf{S})-\sum_{j=0}^{r^*_C}U_j)^2}{\sum_{j=r^*_C+1}^{p}U_j}, \tag{1}$$

and the quantized continuous part $\chi^{2*}_Q(\mathbf{T})$ takes the form:

$$\chi^{2*}_Q(\mathbf{T})=\sum_{j=1}^{r^*_Q}\frac{(DV^{[j]}_Q(\mathbf{T})-V_j)^2}{V_j}+\frac{(\sum_{j=1}^{r^*_Q}DV^{[j]}_Q(\mathbf{T})-\sum_{j=1}^{r^*_Q}V_j)^2}{\sum_{j=r^*_Q+1}^{q}V_j},$$

where p and q are the numbers of attributes from categorical and continuous data, respectively.

If the detected pattern passes a statistical test, we then proceed to extract a cluster by determining the cluster center $\boldsymbol{C}$ and estimating cluster radius $\boldsymbol{R}$ for mixed data. Therefore, a cluster center $\mathbf{C}$ is chosen where the χ^2_w has the maximum value. It is chosen to be:

$$\mathbf{C}=\underset{(\mathbf{S},\mathbf{T})}{\arg\max}\,\chi^2_w.$$

Zhang et al. (2006) [8] gave the definition of radius which is the maximum distance of the data points in this cluster to its center. Radius is the distance at which the DV has its very first local minimum. Therefore, it is defined categorical Radius $R_C(\mathbf{C})$ as:

$$R_C(\mathbf{C})=\min_{0<j<p_C}\{j|DV^{[j]}_C(\mathbf{C})<\text{mim}(DV^{[j-1]}_C(\mathbf{C}),DV^{[j+1]}_C(\mathbf{C}))\}-1.$$

For the quantized continuous part of the data, the optimal cut-off point is used as the quantized continuous radius $R_Q(\mathbf{C})$.

The step-by-step guide to our method is

Step 1. For each position S, we calculate HD in the categorical data; further, we obtain DV_C.

Step 2. Standardize the continuous data and quantize the standardized data at a selected level. For each position calculate Hamming distance for quantized continuous data to obtain DV_Q.

Step 3. Compare DV_C, DV_Q with corresponding expected values UDV_C and UDV_Q.

Step 4. Determine cut-off points $r_C^*(\mathbf{S})$ and $r_Q^*(\mathbf{T})$ for categorical and quantized continuous data respectively; and further calculate the corresponding modified Chi-squared statistic $\chi_C^{2*}(\mathbf{S})$ and $\chi_Q^{2*}(\mathbf{T})$ and obtain the weighted local chi-square test statistic

$$\chi_w^{2*}(\mathbf{S},\mathbf{T}) = \frac{q}{p+q}\chi_C^{2*}(\mathbf{S}) + \frac{p}{p+q}\chi_Q^{2*}(\mathbf{T}).$$

Step 5. Corresponding to the weighted local Chi-squared test, select the largest test statistic $\chi_w^{2*}(\mathbf{S},\mathbf{T})$; compare it with critical value $\chi_{(0.05)}^{2*}$ at the right tail. If the $\max(\chi_w^{2*}(\mathbf{S},\mathbf{T}))$ is smaller than $\chi_{(0.05)}^{2*}$, stop the algorithm; otherwise, continue to step 6.

Step 6. Assign the position that has the largest test statistic $\chi_w^{2*}(\mathbf{S},\mathbf{T})$ as a center. Categorical data and continuous data share the same center position but with their own data points.

Step 7. Calculate categorical a radius R_C and continuous radius R_Q; label all data points within a radius in the cluster; record corresponding $\chi_C^{2*}(\mathbf{S})$ and $\chi_Q^{2*}(\mathbf{T})$; remove them from the current data set.

Step 8. Repeat Steps 1 to 6 until no more significant clusters are detected.

Step 9. Prune the membership assignment by calculating the minimum distance from each data point to center positions; If the membership is assigned differently to categorical data and continuous data, we further compare their p-values which are calculated from $\chi_C^{2*}(S)$ and $\chi_Q^{2*}(S)$; Re-assign the membership to the one with the larger p-value by the one with the smaller p-value.

Step 10. Compute F-test statistic to choose the best-quantized level and corresponding clustering results as the final results.

4. Results

We conduct simulation studies and real data analysis to examine the performance of our proposed method. Classification rates and information gains are calculated to compare the performance of our proposed method with AutoClass.

4.1. Simulation Studies

In this section, we compare our method with AutoClass under various simulation settings. The simulation results are shown in Tables 1–4. All attributes are generated independently. The simulation setting is as the following:

1. Set the number of categorical attributes $p = 10$ and each attribute takes m_j levels which are randomly selected from the set $\{4, 5, 6\}$; Set the number of continuous attributes $q = 9$.
2. Set the number of clusters $K_C = K_Q = 3$ or $K_C = K_Q = 5$. The 3 cluster centers $\mathbf{C}_k$ are denoted as $C_k = (c_{k,1}, \cdots, c_{k,10})$, $k = 1, \cdots, 3$. The 5 cluster centers $\mathbf{C}_k$ are denoted as $C_k = (c_{k,1}, \cdots, c_{k,10})$, $k = 1, \cdots, 3$. For categorical centers, ensure the Hamming distance between any two of the centers is at least great than 5. For the continuous portion of data, choose a set of cluster means as 2, 8, and 16 for 3 clusters, or 2, 8, 16, 20, and 35 for 5 clusters;
3. Set sample size $N = 200$ with cluster size $n_1 = 130$, $n_2 = 45$, and $n_3 = 25$; or set sample size $N = 1000$ with the cluster size $n_1 = 500$, $n_2 = 200$, $n_3 = 100$, $n_4 = 100$, and $n_5 = 100$; or set sample size $N = 10{,}000$ with the cluster size $n_1 = 5500$, $n_2 = 3000$, $n_3 = 1500$;
4. For categorical data, in the k^{th} cluster with center $\mathbf{C}_k$, generate n_k 10-attributes vectors independently. More specifically, generate for each attribute from a multinomial distribution with a center probability of 0.7 and the rest probabilities are identically equal to $0.3/(m_j - 1)$; For continuous data, n_k 9-attributes vectors are 9 independent normal random variables with $\boldsymbol{\mu} = \mathbf{C}_k$ and σ^2 ranging from 0.25, 0.5 and 1, respectively.

In our numerical results, the average classification rate (CR) and information gain (IG) rate with their corresponding standard deviations are used to evaluate the method's performance. The CR measures the accuracy of an algorithm to assign data points to correct clusters. With given K clusters, the CR is defined by

$$CR(K) = \sum_{k=1}^{K} \frac{\tilde{n}_k}{n},$$

where n is the total number of data points and $\tilde{n}_k$ is the number of data points that have been correctly assigned to cluster k by an algorithm. Obviously, $0 \leq CR(K) \leq 1$, and a larger $CR(K)$ value indicates better performance of clustering. The information gain is an alternative criterion for assessing the performance of the clustering algorithm. It is the so-called cluster purity proposed by Bradley et al. (1998) [3]. Cluster purity essentially measures the information gain, which is the difference between the total entropy and weighted entropy for a given data partition, namely

$$information\ gain(IG(K)) = total\ entropy - weighted\ entropy(K),$$

where the weighted entropy is calculated by

$$weighted\ entropy(K) = \sum_{k=1}^{K} \frac{n_k}{n} \times cluster\ entropy(k),$$

with

$$cluster\ entropy = -\sum_{l=1}^{L} \frac{\tilde{n}_l^k}{n_k} \log_2 \left\{ \frac{\tilde{n}_l^k}{n_k} \right\},$$

where $\tilde{n}_l^k$ is the number of data points with true label l in cluster k, n_k is the number of data points known in cluster k, and L is the known number of classes. In this chapter, we take a ratio of IG(K)/total entropy, named information gain rate (IGR), which is similar to the classification rate between 0 to 1. It is necessary to point out that in some situations, the information gained may lead to misleading. For example, in our simulation studies, IG may be equal to 1 which means perfect clustering. However, indeed, it splits each true cluster into two clusters which is a wrong classification. This misleading situation happens in Tables 1 and 2.

Table 1 shows the selection of quantization levels for a continuous portion of the data. As mentioned in Section 2.2, we use the largest F values to choose the selected quantization level which gives the best classification rate. Tables 2–4 provide results from simulated data with various settings of different sample sizes, number of clusters, and cluster sizes. The number of replications is 500 for Tables 2 and 3, and 100 for Table 4. Table 2 is obtained by analyzing simulated data with a sample size of 200 with 3 clusters of the sizes of 130, 45, and 25. Simulated data for Table 3 has a sample size of 1000 and the number of clusters is 5, and each cluster size is 500, 200, 100, 100, and 100, respectively. Table 4 provides results from simulated data having a sample size of 10,000 with 3 clusters and each cluster size of 5500, 3000, and 1500, respectively.

As shown by Tables 2–4, our proposed algorithm consistently has a higher classification rate in comparison with that from AutoClass in all three different settings. For the three chosen settings, the mean classification rates, and information gain rates of the two algorithms are getting closer to each other and could even be identical. Table 3 shows us that our algorithm has higher IG rates compared to AutoClass. In Tables 2 and 4, our algorithm has IG rates varying from 0.8923 to 0.93333. Although AutoClass could achieve one in some cases, this does not imply a perfect clustering because AutoClass tends to split each true cluster into unnecessary more clusters. Hence, overall, all tables show us that our algorithm has better performance in terms of CR and IGR by comparing it to AutoClass.

Table 1. Quantization levels. The means of F statistics, CR, and IG are obtained based on 500 replications.

Discretized Levels	Mean (F)	Mean (CR)	Mean (IGR)
5	630.1573	0.8302	0.7130
6	1523.4557	0.8455	0.7667
7	1722.3260	0.8227	0.6960
8	3223.9477	0.8635	0.7729
9	3916.3388	0.8816	0.7958
10	3708.5293	0.8682	0.7689
11	6444.7055	0.9085	0.8573
12	4778.9851	0.8893	0.8114
13	4912.8477	0.8907	0.8116
14	4262.3990	0.8907	0.8135
15	4000.3948	0.8879	0.8095
16	4234.9993	0.8863	0.7992
17	3549.8632	0.8787	0.7853
18	4042.0805	0.8785	0.7833
19	3657.4556	0.8768	0.7785
20	4303.8698	0.8872	0.8010

Table 2. Average CR and IGR with corresponding standard deviation for each method based on the simulated data of sample size 200 with 3 clusters; each cluster has sizes 130, 45, and 25, respectively. The mean values for each cluster are 2, 8, and 16 respectively. With the same set of means, the different variances, 0.25, 0.5, and 1 are compared. The number of replications is 500.

	AutoClass (Var = 0.25)	Ours (Var = 0.25)	AutoClass (Var = 0.5)	Ours (Var = 0.5)	AutoClass (Var = 1)	Ours (Var = 1)
CR Mean	0.6424	0.9556	0.6335	0.9292	0.6325	0.9370
CR Std	0.0021	0.0035	0.0015	0.0069	0.0015	0.0060
IGR Mean	1.0000	0.8923	1.0000	0.9085	1.0000	0.9148
IGR Std	<0.0001	0.0148	<0.0001	0.0094	<0.0001	0.0070

Table 3. Average CR and IGR with corresponding standard deviation for each method based on the simulated data of the sample size 1000 with 5 clusters; each cluster has sizes 500, 200, 100, 100, and 100, respectively. The mean values for each cluster are 2, 8, 16,20, and 35, respectively. With the same set of means, the different variances, 0.25, 0.5, and 1 are compared. The number of replications is 500.

	AutoClass (Var = 0.25)	Our (Var = 0.25)	AutoClass (Var = 0.5)	Ours (Var = 0.5)	AutoClass (Var = 1)	Ours (Var = 1)
CR Mean	0.5638	0.8747	0.5598	0.8792	0.5615	0.8777
CR Std	0.0016	0.0185	0.0015	0.0179	0.0014	0.0189
IGR Mean	0.7337	0.9228	0.7338	0.9174	0.7338	0.9235
IGR Std	<0.0001	0.0021	<0.0001	0.0049	<0.0001	0.0037

Table 4. Average CR and IGR with corresponding standard deviation for each method based on the simulated data of sample size 10,000 with 3 clusters; each cluster has sizes 5500, 3000, and 1500, respectively. Continuous data are from a multivariate t-distribution with degree freedom 5, 15, and 30, respectively. With the same set of means, the different variances, 0.25, 0.5, and 1 are compared. The number of replications is 100.

	AutoClass (Var = 0.25)	Our (Var = 0.25)	AutoClass (Var = 0.5)	Ours (Var = 0.5)	AutoClass (Var = 1)	Ours (Var = 1)
CR Mean	0.8120	0.9689	0.8231	0.9689	0.8202	0.9641
CR Std	0.0019	0.0031	0.0023	0.0031	0.0033	0.0034
IGR Mean	1.0000	0.9333	1.0000	0.9333	1.0000	0.9323
IGR Std	<0.0001	0.0067	<0.0001	0.0067	<0.0001	0.0048

4.2. Real Data Analysis

We apply our method on to three real data sets. The first two data sets can be downloaded from the Machine Learning Repository website. One is Heart Data Set and the other one is the Australian Credit Approval Data Set. The third data set is collected by the RAND center at the University of Michigan.

Heart Data and Australian Credit Approval Data are downloaded from the Machine Learning Depository at the University of California at Irvine. Heart data contains 7 categorical, 6 continuous attributes, and 270 observations. The data provided the memberships for each observation. There are 2 clusters, absence, and presence. The cluster sizes are 120 and 150, respectively. In the Australian Credit Approval Data Set, there are 8 categorical attributes and 6 continuous attributes. The data set contains two clusters positive or negative with the corresponding cluster sizes 307 and 383. We compared our method with AutoClass. Table 5 shows the results from these two real data sets. From the table, we can tell that our method correctly identified the number of clusters for both data sets, while, AutoClass could not detect correct cluster numbers. In addition, our method has a higher classification rate compared to AutoClass. Our method has a classification rate of 81.48% for Heart data and 73.62% for Credit data. However, AutoClass has 44.44% and 52.71%.

Table 5. Two Real Data Results from two comparison methods. Heart data have 2 clusters with a sample size of 270 and Australian data has 2 clusters with a sample size of 690.

	Heart		**Australian**	
	AutoClass	**Ours**	**AutoClass**	**Ours**
CR	0.4444	0.8148	0.5217	0.7362
IGR	0.2754	0.6975	0.2761	0.8314
Number of clusters	5	2	7	2

We apply our proposed method to the health and retirement study (HRS) data set. Information about health, financial situation, family structure, and health factors was collected by the RAND center at the University of Michigan. We focus on the analysis of the status of depression depicted in the data set. Depression among children and adolescents is common but frequently unrecognized. The clinical spectrum of depression can range from simple sadness to major depressive disorders. A depression diagnosis is often difficult to make because clinical depression can manifest in so many different ways. Observable or behavioral symptoms of clinical depression may be minimal despite a person's mental turmoil. The general population can then be partitioned naturally into two groups: depressed individuals and not depression people. We choose this scenario as the third test case for our clustering algorithm and compare its performance of ours with AutoClass.

We perform clustering based on six health factors: Smoking, Restless Sleep, High Blood Pressure, Frequent Vigorous Physical Activity, Difficulty in Walking, and Age (in months). Depression status is recorded as a binary response variable with $16,250$ depressed and $2,608$ non-depressed individuals; Categorical variables, Smoking, and Restless-sleep, take binary values; Difficulty-in-Walking, takes values $0, 1, 2$, or 9; Frequent Vigorous Physical Activity has values $1, 2, 3, 4$, or 5; High Blood Pressure takes values $0, 1, 3$, or 4; continuous variable, Age(in month), has a range from 224 to $1,232$ with a mean value of 801. For each individual, we include only for which all of the factors were recorded. In total, there are $18,858$ people included in the analysis. Our clustering method correctly identified two clusters. AutoClass, however, detects nine clusters. Tables 6 and 7 report the confusion matrix obtained by our method and AutoClass, respectively. In the Non-depressed group, our method correctly detected 86.75% of non-depressed individuals. In the depressed group, 30.98% of individuals are correctly detected. Since AutoClass finds 9 clusters, it is not feasible to make a fair comparison. Therefore, we describe the nine clusters declared by the AutoClass for the sake of completeness. Table 7 listed AutoClass clustered nine groups and the number of true depression and non-depression patients in each group. Since the

depressed group is much smaller than the non-depression group, the information gain is a not suitable measure since the percentage of the depressed group is always small in comparison with the non-pressed group. The information gain for both our method and AutoClass is small and deemed not informative.

Table 6. Confusion Matrix for our method.

		Our Method		
		Non-Depressed	**Depressed**	**Total**
True	Non-depressed	14,097	2153	16,250
	Depressed	1800	808	2608
	Total	15,897	2961	18,858

Table 7. Confusion Matrix for AutoClass.

		AutoClass									
		Clst1	**Clst2**	**Clst3**	**Clst4**	**Clst5**	**Clst6**	**Clst7**	**Clst8**	**Clst9**	**Total**
True	Non-depressed	3117	2362	1457	2039	1749	2032	1915	1201	378	16,250
	Depressed	216	217	781	335	461	158	144	223	73	2608
	Total	3333	2579	2238	2374	2210	2190	2059	1424	451	18,858

5. Discussion

We have proposed a clustering method that uses statistical distances and tests. Numerical results show that the proposed method outperforms the AutoClass algorithm based on classification rate and entropy measure. The proposed method does not em- ploy a global distance function or a parametric model. For future work, we could consider extending the proposed method to cluster spatial and temporal data.

6. Conclusions

Mixed data are prolific in scientific research such as business, engineering, life sciences, etc. It is imperative to develop a method that can cluster mixed data in order to discover true and significant underlying structures of a data set and classify observations into different subsets. We propose a non-parametric method that uses a local weighted chi-squared statistic to determine underlying clusters. The proposed algorithm does not require any model assumption for attributes or any expensive numerical optimization procedures. Because the proposed algorithm extracts clusters sequentially with one cluster at each iteration, it does not need any convergence criterion. The algorithm is terminated when all data points have been used and no more cluster centers can be detected. Consequently, our algorithm automatically produces the number of clusters, and the resulting partition is unique. When compared with the benchmark clustering algorithm for mixed data, AutoClass, we find that our algorithm outperforms AutoClass in various settings and produces similar accuracy in other settings.

Author Contributions: Conceptualization, X.W., X.G. and Y.X.; methodology, X.W., X.G. and Y.X.; formal analysis, Y.X. and X.W.; writing—original draft preparation, Y.X.; writing—review and editing, X.W. and X.G.; supervision, X.W. and X.G. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable for studies not involving humans or animals.

Data Availability Statement: 1. Heart disease data: https://archive.ics.uci.edu/ml/datasets/Heart+Disease (accessed on 19 November 2022). 2. Australian Credit Approval: https://archive.ics.uci.edu/ml/datasets/statlog+(australian+credit+approval) (accessed on 19 November 2022). 3. RAND HRS

data (Version O): https://hrsdata.isr.umich.edu/data-products/rand-hrs-archived-data-products (accessed on 19 November 2022).

Acknowledgments: Yawen X. acknowledged Xiaogang W.'s and Xin G.'s supervision and support.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

HD	Hamming Distance
DV	Distance Vector
UDV	Uniform Distance Vctor
CR	Classification Rate
IG	Informtion Gain
IGR	Inofrmaiton Gain Rate

References

1. Kaufman, L.; Rousseeuw, P.J. *Finding Groups in Data: An Introduction to Cluster Analysis*; John Wiley & Sons: New York, NY, USA, 2009.
2. Banfield, J.D.; Raftery, A.E. Model-based Gaussian and non-Gaussian clustering. *Biometrics* **1993**, *49*, 803–821. [CrossRef]
3. Bradley, P.S.; Fayyad, U.M.; Reina, C.A. *Scaling EM (Expectation-Maximization) Clustering to Large Databases*; Microsoft Research: Redmond, WA, USA, 1998; pp. 0–25.
4. Fraley, C.; Raftery, A.E. How Many Clusters? Which Clustering Method? Answers Via Model-Based Cluster Analysis. *Comput. J.* **1998**, *41*, 578–588. [CrossRef]
5. Huang, Z. Extensions to the k-means algorithm for clustering large data sets with categorical values. *Data Min. Knowl. Discov.* **1998**, *2*, 283–304. [CrossRef]
6. Cheeseman, P.C.; Stutz, J.C. Bayesian classification (AutoClass): Theory and results. *Adv. Knowl. Discov. Data Min.* **1996**, *180*, 153–180.
7. Ahmad, A.; Dey, L. A K-mean clustering algorithm for mixed numeric and categorical data. *Data Knowl. Eng.* **2007**, *63*, 503–527. [CrossRef]
8. Zhang, P.; Wang, X.; Song, P.X.-K. Clustering Categorical Data Based on Distance Vectors. *J. Am. Stat. Assoc.* **2006**, *101*, 355–367. Available online: http://www.jstor.org/stable/30047463 (accessed on 19 November 2022). [CrossRef]
9. Gersho, A.; Gray, R.M. *Vector Quantization and Signal Compression*; Springer: Berlin, Germany, 1992; pp. 407–485.
10. Graf, S.; Luschgy, H. *Foundations of Quantization for Probability Distributions*; Springer: Berlin/Heidelberg, Germany, 2000.
11. Roman, S. *Coding and Information Theory*; Springer Science & Business Media: New York, NY, USA, 1992; p. 134.
12. Laboulais, C.; Ouali, M.; Le Bret, M.; Gabarro-Arpa, J. Hamming distance geometry of a protein conformational space: Application to the clustering of a 4-ns molecular dynamics trajectory of the HIV-1 integrase catalytic core. Proteins. *Data Knowl. Eng.* **2002**, *47*, 169–179. [CrossRef] [PubMed]