CONNECTIVITY-BASED CLUSTERING FOR MIXED DISCRETE AND CONTINUOUS DATA

Mahfuza Khatun[1] and Sikandar Siddiqui[2]

[1]Jahangirnagar University, Savar, 1342 Dhaka, Bangladesh
[2] Deloitte Audit Analytics GmbH, Europa-Allee 91,
60486 Frankfurt, Germany

ABSTRACT

This paper introduces a density-based clustering procedure for datasets with variables of mixed type. The proposed procedure, which is closely related to the concept of shared neighbourhoods, works particularly well in cases where the individual clusters differ greatly in terms of the average pairwise distance of the associated objects. Using a number of concrete examples, it is shown that the proposed clustering algorithm succeeds in allowing the identification of subgroups of objects with statistically significant distributional characteristics.

KEYWORDS

Cluster analysis, mixed data, distance measures

1. INTRODUCTION

In the field of applied statistics, “clustering” and “cluster analysis” are collective terms for all procedures by which individual objects can be aggregated into groups of mutually similar entities.

One set of methods frequently used to perform this task, commonly referred to as centroid-based approaches, model a cluster as a group of entities scattered around a common central point. However, this may be counter-intuitive since many observers would also tend to group entities together if they are scattered along a common, not necessarily linear path or surface, rather than a common central point.

In response to this challenge, another class of clustering methods, often summarized under the general term of “connectivity-“ or “density-based” approaches (Kriegel et al. [13]), has been developed. In this context, clusters are defined as groups of entities located in regions of high density and separated by near-empty areas within the sample space. Widely used examples of such connectivity-based clustering procedures are DBSCAN (Ester et al., [5]) and OPTICS (Ankerst et al., [2]); see also Oduntan [15].

The current paper addresses a key challenge that may occur in this context: It consists of the possibility that the variables in the relevant datasets may be of mixed type, i.e. some of them may be continuous, while others may either be ordered and discrete (like in the case of school grades) or unordered and discrete (like in the case of country identifiers). This raises the question of how the pairwise distances between the individual entities, which are key inputs for separating low- and high-density regions, are to be measured. A possible solution to this problem is proposed in

 DOI:10.5121/ijci.2024.130303

Section 2 of this paper. The proposed clustering procedure itself, which is closely related to the concept of shared neighbourhoods presented in Houle et al. [10], is described in Section 3. Section 4 summarizes the outcomes of a number of exemplary applications. Section 5 concludes.

2. MEASURING DISTANCES BETWEEN ENTITIES

2.1. Starting Point

As in Khatun and Siddiqui [12], we assume that there is a dataset consisting of N entities $i = 1, \ldots, N$, each of which is characterised by a tuple $q_i = \{ x_i, v_i, z_i \}$ of features.

- x_i is a realisation of a $(K_1 \times 1)$ column vector X of numerical variables that either are continuous or treated as continuous for practical reasons
- v_i is a realisation of a $(K_2 \times 1)$ vector V of ordered discrete variables, beginning with 1 numbered consecutively in steps of 1, and
- z_i is a realisation of a $(K_3 \times 1)$ vector Z of unordered, discrete variables.

2.2. Pairwise Distance with Respect to Continuous Variables

The distance between two entities i and j with respect to the values of X is measured by the Manhattan Distance (see, e.g., Yang [20]) between the standardized x_i and x_j values as follows:

$$d_x(i,j) = \sum_{s=1}^{K_1} \frac{|x_{i,s} - x_{j,s}|}{r(X_s)} \quad (1)$$

where $r(X_s)$ denotes the range of the observed values of X_s, i.e. the difference between the sample maximum and the sample minimum.

In this context, the Manhattan Distance is preferred to the more commonly used Euclidean distance because when using the former, the contrast between the distances from different data points shrinks less rapidly as the dimension K_1 of X grows; see Aggarwal, Hinneburg, and Keim [1]. In addition, the application of the distance measure (1) simplifies the consolidation of distance measures for the different variable types involved, as will become obvious below.

2.3. Pairwise Distance with Respect to the Ordered Discrete Variables

The distance between i and j with respect to the values taken by the components of V can be measured by

$$d_v(i,j) = \sum_{s=1}^{K_2} \frac{|v_{i,s} - v_{j,s}|}{m_s} \quad (2)$$

where m_s is the number of possible realisations of the s-th ordered discrete variable.

This way of proceeding is justified as follows: If V_s is an ordered, discrete variable ranging from 1 to m_s in steps of 1, then the actual value $v_{i,s}$ can be assumed to be dependent on the value taken by a latent (=unobservable) variable $v_{i,s}^* \in]0; 1]$ as follows:

$$v_{i,s} = j \text{ if } v_{i,s}^* \in](j-1)/ m_s;\ j/m_s\ ] \quad (3)$$

The distance between the mid-points of two neighbouring sub-intervals given in (3) equals $1/m_s$.

2.4. Pairwise Distance based on Unordered Discrete Characteristics

With regard to *Z*, the distance between two entities *i* and *j* can be quantified by their separateness or lack of overlap (see, e.g., Stanfill and Waltz [16]), based on the Hamming Distance (Hamming [9]):

$$d_z(i,j) = \sum_{s=1}^{K_3} \frac{I(z_{i,s} \neq z_{j,s})}{m_s} \tag{4}$$

In the above equation, $I(\cdot)$ is the indicator function that equals 1 if the condition in brackets is fulfilled and 0 if not. The scalar m_s stands for the number of distinct possible realisations of the *s*-th unordered discrete variable.

The distance prevailing between two specific realisations $z_{i,s}$ and $z_{j,s}$ of the s-th unordered discrete variable Z_s with respect to the values thus equals 0 if $z_{i,s}$ equals $z_{j,s}$, and $1/m_s$ if not. The distance between pairs of observations with different values of Z_s, as measured by (3), thus shrinks as the number of possible realisations of the relevant variable increases. This normalisation rule is motivated by the idea that the finer the classification scheme according to which individual entities are grouped, the smaller the average number of entities per group, and the less certain we can be that differences in group membership reflect actual disparities between the entities, rather than merely random "noise".

2.5. Overall Pairwise Distance Between Two Entities

The overall distance score between two entities *i* and *j* can then be calculated by summing up the distance measures for the different types of variables given in (1), (2), and (4).

$$d(i,j) := d_x(i,j) + d_v(i,j) + d_z(i,j) \tag{5}$$

3. CLUSTERING PROCEDURE

As mentioned in the introduction, the clustering procedure proposed in this section is based on the concept of shared neighbourhood, as presented in the seminal paper by Houle et al. [10]. This approach is adopted here because, according to the authors, it is less affected by the "curse of dimensionality" (Bellman [3]). What is meant by this term is that as the number of variables under consideration grows, the size differences between the pairwise distances of the individual data points decrease, which makes it increasingly impossible to form meaningful clusters based on such distances.

In line with the above specifications, it is possible to calculate, for each entity *i* in the sample, the distance between itself and each of the remaining (*N*-1) entities, and to sort the results of these calculations in ascending order. Let $\delta_i(1) \leq \delta_i(2) \leq \ldots \leq \delta_i(N\text{-}1)$ denote the sorted distances from *i*, and let $g > 0$ be a user-specified integer number. Then, the adjacency set S_i of entity *i*, is defined as the sets of all entities $j \neq i$ for which the inequality $d(i,j) \leq \delta_i(g)$ holds:

$$S_i := \{ j \neq i \mid d(i,j) \leq \delta_i(g) \} \tag{6}$$

Two entities j and i are considered mutually interlinked if their adjacency sets S_i and S_j have at least one element in common:

$$S_i \Leftrightarrow S_j \quad \Leftrightarrow \quad S_i \cap S_j \neq \{\varnothing\} \tag{7}$$

The proposed clustering procedure can then be summarized as follows:

Listing 1: Clustering Procedure

Step 1:	Gather all entities in the sample in the subset of hitherto unassigned entities
Step 2:	Initialize the cluster index c as 0.
Step 3:	Increase the cluster index c by 1.
Step 4:	Set the number of elements in cluster c, denoted by n_c, to 0
Step 5:	Set i to 1.
Step 6:	If • entity number i has not yet been assigned to a cluster and • n_c exceeds 0 and • entity number i and at least one element of cluster c are mutually interlinked then • assign entity number i to cluster c, • remove entity number i from the subset of hitherto unassigned entities, and • increase n_c by 1
Step 7:	If • entity number i has not yet been assigned to a cluster, and • n_c equals 0, then • assign entity number i to cluster c, • remove entity number i from the subset of hitherto unassigned entities, and • increase n_c by 1
Step 8:	Increase i by 1
Step 9	If $i \leq N$, continue with Step 6
Step 10	If $i > N$, and if there is at least one object in the subset of hitherto unassigned entities that is mutually interlinked with at least one element of cluster c, continue with Step 5.
Step 11	If $i > N$ and if there is not a single element in the subset of hitherto unassigned entities is mutually interlinked with at least one element of cluster c, continue with Step 3.
Step 12	If $i > N$ and the set of hitherto unassigned entities is empty, terminate.

An implementation of this procedure in the matrix language GAUSS, together with the datasets for the examples from the following section, is available on the WWW via https://drive.google.com/drive/folders/1putjdHHMJjg2TafWenFl9lMBr3ShxC93?usp=drive_link .

The above procedure will cause each of the entities in the sample to be unequivocally assigned to a single cluster. However, particularly when the number g of neighbouring entities from which the adjacency set of each entity i is derived is small (say, e.g. 1 or 2), some or even many objects will be assigned to "degenerate" clusters comprising only a single object. On the other hand, raising the value of g above a certain threshold (the level of which depends on the distributional characteristics of the underlying dataset) will cause all objects to be gathered in a single, maximally heterogeneous group. It thus becomes obvious that the choice of g implies a trade-off between the potentially conflicting objectives of within-cluster homogeneity on one hand and

inclusiveness (i.e. the assignment of as many entities as possible to valid, or "non-degenerate", clusters) on the other.

The proposed solution to this problem is to compare different possible outcomes of the clustering procedure with a measure of separation accuracy that can be calculated as follows: Let ξ_{min} denote a user-defined minimum cluster size, $c(i)$ the index number of the cluster to which object i has been assigned, and $n_{c(i)}$ the number of objects in that cluster. Then, the quantity

$$d_i^{\circ} := \begin{cases} \min\limits_{j \neq i;\ j \in c(i)} d(i,j) & if\ n_{c(i)} \geq \xi_{min} \\ \min\limits_{j \neq i} d(i,j) & if\ n_{c(i)} < \xi_{min} \end{cases} \tag{8}$$

equals the distance between object i and its closest neighbour within its cluster, provided that $c(i)$ at least reaches the specified minimum size, whereas

$$d_i^{\circ\circ} := \begin{cases} \min\limits_{j \neq i;\ j \notin c(i)} d(i,j) & if\ n_{c(i)} \geq \xi_{min} \\ \min\limits_{j \neq i} d(i,j) & if\ n_{c(i)} < \xi_{min} \end{cases} \tag{9}$$

equals the distance between object i and its closest neighbour outside its cluster whenever $c(i)$ does not fall short of ξ_{min}. Then, for any given value ξ_{min}, the particular value g^* of g that maximizes the quantity

$$\vartheta(\xi_{min}, g) := (1/\mathrm{N}) \sum_{i=1}^{N} (1 + d_i^{\circ\circ})/(1 + d_i^{\circ}) \tag{10}$$

can be looked upon as the one that maximizes the separation accuracy associated with the chosen value of ξ_{min}.

4. EXEMPLARY APPLICATIONS

4.1. Two-Dimensional Datasets with Continuous Variables Only

Although the main purpose of the proposed method is to allow the clustering of mixed data, some of its key characteristics are probably best understood when applying it to a small set of deliberately simple cases. The first important feature is that, unlike the centroid-based procedures, the procedure from Sections 2 and 3 can identify clusters of arbitrary shape rather than being limited to ones with round or oval profiles. In an exemplary manner, this is shown for the "Aggregation" dataset examined by Gionis, Mannila and Tsaparas (2007). Figure 1 displays the outcome of an application of the proposed procedure in this case:

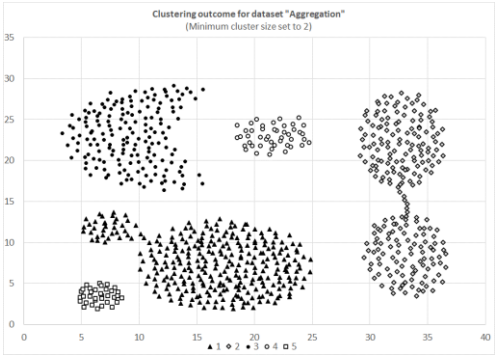

Figure 1

In the proposed procedure, the measure of neighbourliness that is used to decide whether two or more data points belong together or are considered separate is not defined by a fixed distance threshold; rather it is based on the presence or absence of at least one common neighbour. In a setup where the individual clusters differ greatly in terms of the average pairwise distance of the associated objects, this feature enables the procedure from Section 3 to identify such accumulations of objects nevertheless, as is shown by its application to the “toy dataset” dealt with in Jain and Law [11]:

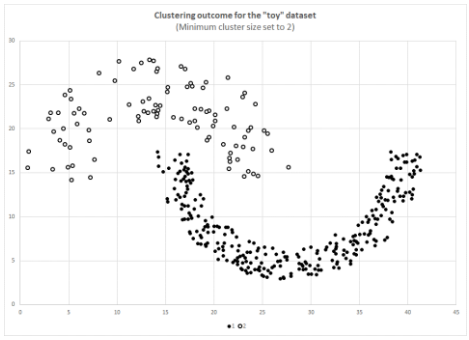

Figure 2

However, the application of the proposed method to the R15” dataset studied in Veenman, Reinders and Backer [18] also points to the flip side of the apparent advantages of our algorithm: Because of its emphasis on common close neighbours, it tends to merge groups of objects surrounding two or more different central points into a single group whenever there is some degree of overlap between them. This can sometimes lead to counterintuitive results, as Figure 3 indicates: Here, our algorithm forms a single cluster out of eight centrally located point clouds near the centre of the diagram, although most human observers would probably have perceived them as separate groups.

The outcome for the dataset “Unbalance” examined by Rezaei and Fränti [14] (see Figure 4) further underscores this point. Here, most human observers would probably have split Cluster 1 in four and Cluster 2 into three separate “sub-clusters”.

The above findings show that the proposed approach is far from a universal, objective solution to clustering problems. It should rather be considered one out of several related approaches which, in view of the variety of possible data and research objectives, can produce results with very different degrees of plausibility from case to case.

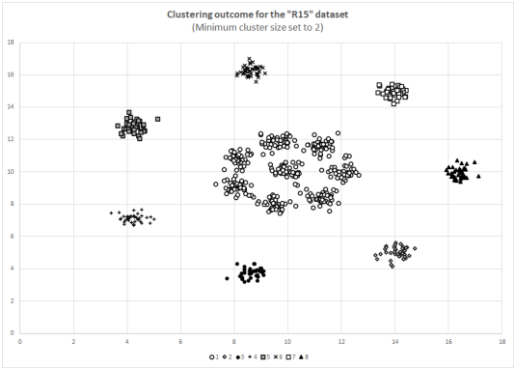

Figure 3

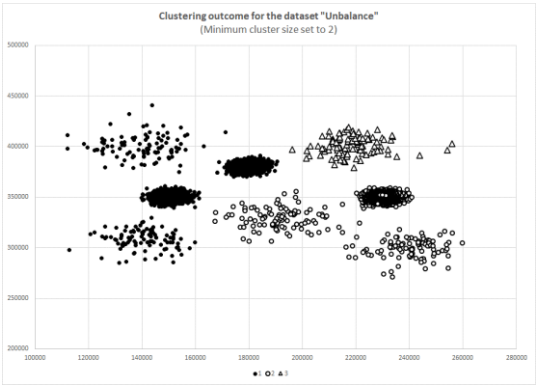

Figure 4

The four examples in this subsection relate to synthetically generated data points from a bivariate distribution involving continuous variables only. The purpose of the two following subsections is to demonstrate that the proposed algorithm can also produce plausible results in real-world situations involving more variables, and variables of different types.

4.2. Country Grouping by Macroeconomic Indicators

The above procedure can be applied to form groups of countries based on similarity comparisons with regard to their geographic location as well as a number of macroeconomic indicators. In the particular case examined here, we use the seven World Bank [19] regions as geographical assignment indicators and a set of four macroeconomic variables, which include (i) GDP per capita, as well as (ii) government debt, (iii) the current account balance, and (iv) the government budget balance, the last three of which are being expressed as a percentage of GDP. This choice is motivated because the four variables just mentioned are among the most commonly used indicators used to assess the resilience of countries to adverse economic shocks; see, e.g., Briguglio et al. [4] for a more comprehensive treatment of this issue. The common source for all the data in use is Trading Economics [17], and the reference date is generally the year-end of 2022. If no data is available for this date, the most recent earlier key date has been chosen. After removing sovereign states with missing data, we end up with a sample of 163 countries.

With the minimum cluster size $\xi_{\min}$ set to 2, application of the above procedure to the dataset compiled accordingly yields an optimum number of close neighbours (g^*) of 2 and leads to two large clusters, one more small cluster of only two countries (Cambodia and the Maldives), and a total of six “outliers” (Afghanistan, Bhutan, Cyprus, Guinea, Indonesia, and Lebanon) that cannot

be assigned to a cluster that reaches or exceeds the above size threshold. Table 4.1.1. enumerates the countries assigned to Cluster 1 and 2 by region. Descriptive statistics of the four continuous variables in use are given in Tables 4.1.2 to 4.1.5.

Table 1: Distribution of countries across Clusters 1 and 2

	Cluster 1	Cluster 2
East Asia & Pacific	Fiji, Japan	Australia, Brunei, China, Laos, Malaysia, Mongolia Myanmar, New Zealand, Papua New Guinea, Philippines, Singapore, South Korea, Thailand, Vietnam
Europe & Central Asia	Albania, Armenia, Belgium, Bosnia and Herzegovina, France, Georgia, Greece, Italy, Macedonia, Moldova, Montenegro, Portugal, Serbia, Spain, Ukraine, United Kingdom, Uzbekistan	Austria, Azerbaijan, Belarus, Bulgaria, Czech Republic, Croatia, Denmark, Estonia, Finland, Germany, Hungary, Kazakhstan, Kosovo, Kyrgyzstan, Iceland, Latvia, Lithuania, Luxembourg, Netherlands, Norway, Poland, Romania, Russia, Slovakia, Slovenia, Sweden, Switzerland, Tajikistan, Turkey, Turkmenistan
Latin America & Caribbean	Argentina, Bahamas, Belize, Bolivia, Brazil, Chile, Colombia, Costa Rica, Dominican Republic, Ecuador, El Salvador, Guatemala, Haiti, Honduras, Jamaica, Mexico, Nicaragua, Panama, Paraguay, Peru, Suriname, Trinidad and Tobago, Uruguay,	Cayman Islands Guyana
Middle East & North Africa	Algeria, Bahrain, Djibouti, Egypt, Iran, Iraq, Jordan, Libya, Palestine, Tunisia	Israel, Kuwait, Malta, Saudi Arabia, United Arab Emirates, Oman, Qatar
North America	Canada, United States	
South Asia	Bangladesh, Nepal, India, Pakistan, Sri Lanka	
Sub-Saharan Africa	Angola, Benin, Botswana, Burkina Faso, Burundi, Cameroon, Cape Verde, Central African Republic, Chad, Comoros, Congo, Ethiopia, Equatorial 1Guinea, Gabon, Gambia, Ghana, Guinea, Guinea Bissau, Ivory Coast, Kenya, Lesotho, Liberia, Madagascar, Malawi, Mauritania, Mauritius, Mozambique, Namibia, Niger, Nigeria, Republic of the Congo, Senegal, Sierra Leone, Rwanda, Seychelles, South Africa, Sudan, Swaziland, Tanzania, Togo, Uganda, Zambia, Zimbabwe	Central African Republic, Liberia, Mauritania, Niger, Senegal, Seychelles

Table 2: Descriptive Statistics by Cluster, GDP per capita

Cluster #	Mean	Std. dev.	Min.	Max	# obs.
1	13829.40	13347.69	705.00	63670.00	95
2	36308.82	26802.29	838.00	115683.00	60
3	11560.00	10189.41	4355.00	18765.00	2
Sample	22067,75	22162,15	705.00	115683.00	163

Table 3: Descriptive Statistics by Cluster, Government debt as a percentage of GDP

Cluster #	Mean	Std. dev.	Min.	Max	# obs.
1	71.38	40.39	14.60	264.00	95
2	46.80	24.67	1.90	160.00	60
3	47.35	14.92	36.8	57.9	2
Sample	**62,32**	**37.89**	**1,9**	**264**	**163**

Table 4: Descriptive Statistics by Cluster, Current account balance as a percentage of GDP

Cluster #	Mean	Std. dev.	Min.	Max	# obs.
1	-3.15	6.56	-23.8	19.2	95
2	0.42	12.36	-26.8	30.5	60
3	-21.75	7.28	-26.9	-16.6	2
Sample	**-2.21**	**10.17**	**-33.80**	**30.50**	**163**

Table 5: Government budget balance as a percentage of GDP

Cluster #	Mean	Std. dev.	Min.	Max	# obs.
1	-4.28	5.07	-35.00	6.50	95
2	-2.41	4.35	-19.30	11.60	60
3	-10.80	4.95	-14.30	-7.30	2
Sample	-3.62	4.88	-35.00	11.60	163

It turns out that the entities assigned to Cluster 1, on average, have a lower GDP per capita and higher ratios of government debt and government budget deficits to GDP than those gathered in Cluster 2. In both of these cases, a standard two-sample t-test leads to a rejection of the null hypothesis of equal means on a confidence level exceeding 99%. Moreover, the average government budget balance and the average current account balance, both expressed as a percentage of GDP, are lower in Cluster 1 than in Cluster 2. In both of these cases, the absolute values of the associated t-statistics exceed the 95% critical value for a two-sided test by far.

Clusters 1 and 2 also exhibit considerable differences between the correlation patterns prevailing between the continuous variables involved, only the most striking of which are mentioned in the following:

- In Cluster 1, the sample correlation coefficient between the GDP per capita and the government budget balance per unit of GDP exceeds 0.5 and is statistically significant on a 95% level. In Cluster 2, the same coefficient is below 0.06 and statistically insignificant.

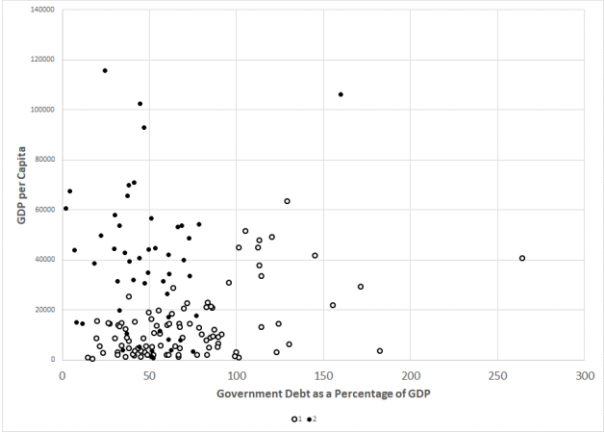

Figure 4:

- In Cluster 1, the correlation coefficient between GDP per capita and the government's budget balance takes a negative value (-0.20), whereas in Cluster 2, the corresponding coefficient has the opposite sign (0.297). A t-test derived a corresponding bivariate linear regression in which the slope parameters were allowed to differ in Cluster 1 and Cluster 2 resulted in the null hypothesis of identical parameter values being rejected on a 99% confidence level.

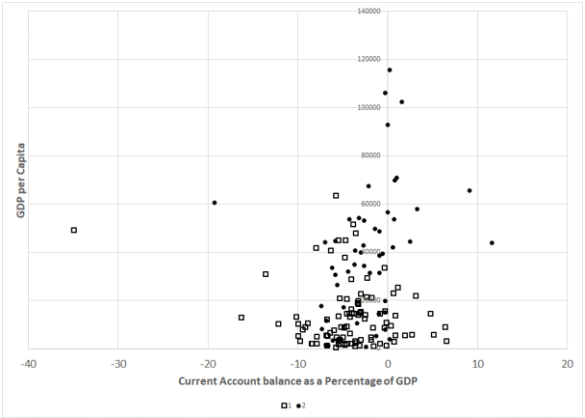

Figure 5:

The above results clearly indicate that, in this particular case, the proposed clustering procedure succeeds in allowing the identification of subgroups of entities with statistically significant distributional characteristics that might not have been detectable with other tools of exploratory data analysis, such as histograms and scatterplots.

4.3. Clusters of Credit Card Applicants

Another dataset to which the procedure from sections 2 and 3 can be applied is the sample of applicants for a specific type of credit card provided in the online complements of Greene's [8] econometrics textbook. The dataset consists of 1,319 observations on 12 variables. The variables used in the current example are listed in Table 4.2.1.

Table 6: Variables Used in the Credit Card Example

Name	**Type**	**Description**
card	Discrete, unordered	= 1 if application was accepted, = 0 otherwise
reports	Treated as continuous	Number of major derogatory reports
age	Continuous	Age of the applicants in years
income	Continuous	Annual income in USD 10 000
owner	Discrete, unordered	= 1 if applicants own their home, = 0 otherwise
selfemp	Discrete, unordered	= 1 if applicants are self-employed, = 0 otherwise
dependents	Treated as continuous	Number of dependents

In this case, the application of the proposed procedure with a minimum cluster size of 2 leads to the formation of two large clusters without any outliers. Cluster 1, the larger of these two, number 1, comprises little more than two-thirds of the sample. Cluster-specific descriptive statistics for those that either are continuous or treated as such are given in Tables 7 to 10 below.

Table 7: Descriptive Statistics by Cluster, Variable "reports"

Cluster #	Mean	Std. dev.	Min.	Max	# obs.
1	0.6212	1.5917	0.0000	14.0000	887
2	0.1180	0.3942	0.0000	3.0000	432
Sample	0.4564	1.3453	0.0000	14.0000	1319

Table 8: Descriptive Statistics by Cluster, Variable "age"

Cluster #	Mean	Std. dev.	Min.	Max	# obs.
1	35.4493	10.1895	0.1667	83.5000	887
2	28.6215	8.3509	0.5000	67.1667	432
Sample	33.2131	10.1428	0.1667	83.5000	1319

Table 9: Descriptive Statistics by Cluster, Variable "income"

Cluster #	Mean	Std. dev.	Min.	Max	# obs.
1	3.6686	1.8618	0.2100	67.1667	887
2	2.7427	1.0347	1.3200	10.9999	432
Sample	3.3654	1.6939	0.2100	13.5000	1319

Table 10: Descriptive Statistics by Cluster, Variable "dependents"

Cluster #	Mean	Std. dev.	Min.	Max	# obs.
1	1.3766	1.3372	0.0000	6.0000	887
2	0.2083	0.4066	0.0000	1.0000	432
Sample	0.9939	1.2477	0.0000	6.0000	1319

The outcome of a two-sample t-test indicates that, for each of the four variables named above, the cluster-specific differences in the means are statistically significant at a 99% confidence level.
Below is information on cluster-specific frequency distributions for the three binary indicator variables in use. Here, too, the absolute values of the t-statistics relating to the differences between the two clusters exceed the critical values for the 99% confidence interval by far.

Table 11:

Cluster #	application accepted	% self-employed	% homeowners
1	66.63%	10.26%	65.50%
2	100.00%	0.00%	0.00%
Sample	77.56%	6.90%	44.04%

In most of the above cases, the cluster-specific pairwise correlation patterns between the four numerical variables involved (“reports”, “age”, “income”, and “dependents”) do not show any pronounced differences. An exception, however, is the relationship between the income variable and the number of dependents, which is positive and statistically significant on a 95% confidence level in the case of Cluster 1 but close to zero in Cluster 2. Hence, the supposition that the proposed clustering method can be applied to identifying subgroups of entities with significantly different statistical characteristics is also confirmed by the results obtained in this case.

4.4. Precautionary Remarks

Given the obvious suitability of the proposed approach in the context of the above examples, it appears necessary to mention a number of problems that cannot be resolved by its application:

- The performance of the algorithm presented and the characteristics of the outcomes obtained are very sensitive to the choice of the minimum cluster size and the number of neighbouring entities from which the adjacency set of each entity is derived. If the latter is chosen too small, the proposed method may result in the formation of "degenerate" clusters with only very few elements.

- Many datasets to which this method can, in principle, be applied may contain one or more “irrelevant” variables that do not contain any information on the basis of which objects can be meaningfully divided into groups of interconnected elements.

- Especially in data sets with many variables, strong dependency relationships between individual attributes, or subsets thereof, can prevent the application of distance-based grouping procedures of the kind described here.

- The problem persists that notions like “distance” or “neighbourhood” become less significant as the dimension of a dataset increases. The method proposed in this paper may help mitigate this under favourable conditions, but it is by no means a complete solution.

- The performance of the proposed algorithm is very sensitive to the structure and distribution of the data to which it is applied. Hence, it remains an open question to what extent the proposed algorithm can be generalized across different datasets.

5. CONCLUSIONS

In this paper, a distance-based clustering procedure for data of mixed type has been proposed. The feasibility of the proposed method and its ability to produce empirically plausible results were demonstrated using some application examples of different nature and complexity. However, the curse of dimensionality and the possible presence of irrelevant or strongly interrelated (groups of) variables remain issues that can lead to great difficulties for such applications. Hence, augmenting the proposed technique with feature selection and/or

dimensionality reduction techniques that may help mitigate this problem is a promising area for future research.

ACKNOWLEDGEMENTS

The authors are grateful for several helpful comments and suggestions by two anonymous referees.

The project underlying this publication was funded by the German Federal Ministry of Economic Affairs and Climate Action under project funding reference number 01MK21002G.

REFERENCES

[1] Aggarwal, C.C., Hinneburg, A., & Keim, D. A. (2001). On the Surprising Behavior of Distance Metrics in High Dimensional Space. In J. van den Bussche, & V. Vianu (Eds.), Database Theory — ICDT 2001. Berlin (Springer).

[2] Ankerst, M., Breunig, M. M., Kriegel, H.-P., & Sander, J. (1999). OPTICS: Ordering Points To Identify the Clustering Structure. In: ACM SIGMOD Record, 28(2), 49-60. https://doi.org/10.1145/304181.304187

[3] Bellman, R. E. (1961). Adaptive Control Processes: a Guided Tour. Princeton (University Press).

[4] Briguglio, L., Cordina, G., Farrugia, N. & Vella, S. (2008). Economic vulnerability and resilience concepts and measurements. Helsinki (United Nations University).

[5] Ester, M., Kriegel, H.P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In E. Simoudis, J. Han & U.M. Fayyad (Eds.), Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96), Washington, D.C. (AAAI Press).

[6] Gionis, A., Mannila, H., & Tsaparas, P. (2007). Clustering aggregation. ACM Transactions on Knowledge Discovery from Data, 1(1), 1-30. https://doi.org/10.1145/1217299.1217303

[7] Gower, J. C. (1971). A general coefficient of similarity and some of its properties. Biometrics, 27, 623-637. https://doi.org/10.2307/2528823

[8] Greene, W.H. (2003). Econometric Analysis. Upper Saddle River, NJ (Prentice Hall).

[9] Hamming, R.W. (1950). Error detecting and error correcting codes. Bell System Technical Journal, 29 (2), 147–160. https://doi.org/10.1002/j.1538-7305.1950.tb00463.x

[10] Houle, M. E., Kriegel, H. P., Kroger, P., Schubert, E., and Zimek, A. (2010). Can shared-neighbor distances defeat the curse of dimensionality? In: M. Gertz, & B. Ludäscher (Eds.), Scientific and Statistical Database Management (SSDBM 2010). Lecture Notes in Computer Science, vol 6187. Berlin and Heidelberg (Springer).

[11] Jain, A. and Law, M. (2005). Data Clustering: A User's Dilemma. In S.K. Pal, S. Bandyopadhyay & S. Biswas (Eds.), Pattern Recognition and Machine Intelligence. (Lecture Notes in Computer Science, vol 3776), Berlin (Springer). https://doi.org/10.1007/11590316_1

[12] Khatun, M., & Siddiqui, S. (2023). Estimating Conditional Event Probabilities with Mixed Regressors: a Weighted Nearest Neighbour Approach. Statistika 103(2), 226-234.

[13] Kriegel, H.-P., Kröger, P. Sander, J., & Zimek, A. (2011). Density-based Clustering. WIREs Data Mining and Knowledge Discovery, 1 (3), 231–240. https://doi.org/10.1002/widm.30

[14] Rezaei, M., & Fränti, P. (2016). Set-matching measures for external cluster validity. IEEE Trans. on Knowledge and Data Engineering 28 (8), 2173-2186. https://doi.org/10.1109/TKDE.2016.2551240

[15] Oduntan, O. I. (2020). Blending Multiple Algorithmic Granular Components: A Recipe for Clustering. Department of Computer Science, The University of Manitoba, Winnipeg, Canada.

[16] Stanfill, C., & Waltz, D. (1986). Toward memory-based reasoning. Commun. ACM, 29(12),1213–1228. https://doi.org/10.1145/7902.7906

[17] Trading Economics (2023): Indicators. Retrieved July 1st, 2023 from https://tradingeconomics.com/indicators .

[18] Veenman, C.J., Reinders, M.J.T., & Backer, E. (2002). A maximum variance cluster algorithm. IEEE Trans. Pattern Analysis and Machine Intelligence, 24(9), 1273-1280. https://doi.org/10.1109/TPAMI.2002.1033218

[19] World Bank (2023). World Bank Units. Retrieved July 1, 2023 from https://www.worldbank.org/en/about/unit

[20] Yang, X.-S. (2019). Introduction to Algorithms for Data Mining and Machine Learning. Amsterdam (Elsevier).