CC BY: Attribution alone 4.0
https://creativecommons.org/licenses/by/4.0/
https://doi.org/10.26219/heal.aueb.5832

ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ
ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

SCHOOL OF INFORMATION SCIENCES & TECHNOLOGY

DEPARTMENT OF STATISTICS

POSTGRADUATE PROGRAM

Distance-Based Methods for Clustering Mixed Type Data. A Review and Comparison Study with Gower's Coefficient.

By

Alexia Elizabeth Hobbs-Ismeris

A THESIS

Submitted to the Department of Statistics

of the Athens University of Economics and Business

in partial fulfilment of the requirements for

the degree of Master of Science in Statistics

Athens, Greece

October 2022

CC BY: Attribution alone 4.0
https://creativecommons.org/licenses/by/4.0/
https://doi.org/10.26219/heal.aueb.5832

CC BY: Attribution alone 4.0
https://creativecommons.org/licenses/by/4.0/
https://doi.org/10.26219/heal.aueb.5832

ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ
ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

ΣΧΟΛΗ ΕΠΙΣΤΗΜΩΝ & ΤΕΧΝΟΛΟΓΙΑΣ ΤΗΣ ΠΛΗΡΟΦΟΡΙΑΣ

ΤΜΗΜΑ ΣΤΑΤΙΣΤΙΚΗΣ

ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ

Μέθοδοι με βάση την αποστάση για Clustering δεδομένων μικτού τύπου. Μια μελέτη ανασκόπησης και σύγκρισης με το συντελεστή Gower.

Αλέξια Ελίζαμπεθ Χόμπς-Ισμέρη

ΔΙΑΤΡΙΒΗ

Που υποβλήθηκε στο Τμήμα Στατιστικής

του Οικονομικού Πανεπιστημίου Αθηνών

ως μέρος των απαιτήσεων για την απόκτηση

Διπλώματος Μεταπτυχιακών Σπουδών στη Στατιστική

Αθήνα

Οκτώβριος 2022

CC BY: Attribution alone 4.0
https://creativecommons.org/licenses/by/4.0/
https://doi.org/10.26219/heal.aueb.5832

CC BY: Attribution alone 4.0
https://creativecommons.org/licenses/by/4.0/
https://doi.org/10.26219/heal.aueb.5832

ACKNOWLEDGEMENTS

I would like to say thank you and give great appreciation to my supervisor, Prof. Ioulia Papageorgiou, for her support and guidance throughout the writing of my thesis. Her extreme kindness, possession of knowledge and continuous availability were of essential help to me throughout this period. I would also like to thank my family, for always supporting me and believing in me.

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

CC BY: Attribution alone 4.0
https://creativecommons.org/licenses/by/4.0/
https://doi.org/10.26219/heal.aueb.5832

CC BY: Attribution alone 4.0
https://creativecommons.org/licenses/by/4.0/
https://doi.org/10.26219/heal.aueb.5832

ABSTRACT

Alexia Elizabeth Hobbs-Ismeris

Distance-Based Methods for Clustering Mixed Type Data. A Review and Comparison Study with Gower's Coefficient.

October 2022

Clustering is a popular approach in data mining applications for discovering, managing, analysing, and extracting critical information from large volumes of data. Clustering algorithms strive to discover homogeneous groups of objects based on attribute values. The majority of clustering algorithms are suitable to either solely numerical or solely categorical data, but not both. However, datasets with mixed data types are common in real life applications and such datasets have been gathered in many fields. In the case of mixed-type variables, one main approach is cluster analysis based on similarity measures. Therefore, creating appropriate similarity measures is a critical step in clustering these datasets. This thesis focuses on similarity measures for mixed data which can be applied in hierarchical cluster analysis, with its main topic being the Gower's similarity measure. Several modifications of the Gower similarity coefficient are examined as well as various additional similarity measures. A simulation study is conducted to assess the performance of the different measures on mixed data while using hierarchical clustering under varying conditions. All the examined similarity measures are compared regarding the quality of the produced clusters by using the Rand index, and therefore assumptions concerning the efficiency of each measure are formulated under different circumstances.

CC BY: Attribution alone 4.0
https://creativecommons.org/licenses/by/4.0/
https://doi.org/10.26219/heal.aueb.5832

CC BY: Attribution alone 4.0
https://creativecommons.org/licenses/by/4.0/
https://doi.org/10.26219/heal.aueb.5832

ΠΕΡΙΛΗΨΗ

Αλέξια Ελίζαμπεθ Χόμπς-Ισμέρη

Μέθοδοι με βάση την απoστάση για Clustering δεδομένων μικτού τύπου. Μια μελέτη ανασκόπησης και σύγκρισης με το συντελεστή Gower.

Οκτώβριος 2022

Η συσταδοποίηση αποτελεί μια δημοφιλής προσέγγιση σε εφαρμογές εξόρυξης δεδομένων από μεγάλους όγκους δεδομένων για την ανακάλυψη, διαχείριση, ανάλυση και εξαγωγή χρήσιμων πληροφοριών. Οι αλγόριθμοι συσταδοποίησης προσπαθούν να ανακαλύψουν ομοιογενείς ομάδες αντικειμένων με βάση τις τιμές των μεταβλητών. Η πλειοψηφία των αλγορίθμων συσταδοποίησης είναι κατάλληλοι είτε για αριθμητικά είτε για κατηγορικά δεδομένα, αλλά όχι για το συνδυασμό και των δύο. Τα δεδομένα που συναντώνται συνήθως στην καθημερινότητα, αποτελούνται ωστόσο απο μεικτούς τύπους δεδομένων. Μία απο τις κύριες προσεγγίσεις για τη συσταδοποίηση μεικτών δεδομένων βασίζεται στη χρήση μέτρων ομοιότητας. Επομένως, η δημιουργία κατάλληλων μέτρων ομοιότητας είναι ένα κρίσιμο βήμα για τη συσταδοποίηση αυτών των συνόλων δεδομένων. Η παρούσα διπλωματική εργασία εστιάζει σε μέτρα ομοιότητας για μεικτά δεδομένα που μπορούν να χρησιμοποιηθούν στην ιεραρχική μέθοδο συσταδοποίησης, με κύριο αντικείμενο μελέτης το μέτρο ομοιότητας του Gower. Συγκεκριμένα, εξετάζονται κάποιες τροποποιήσεις του συντελεστή Gower καθώς και πρόσθετα μέτρα ομοιότητας. Επιπρόσθετα, διεξάγεται μια μελέτη προσομοίωσης για την αξιολόγηση της απόδοσης των διαφορετικών μέτρων σε μεικτά δεδομένα, υπό διαφορετικές συνθήκες. Όλα τα εξεταζόμενα μέτρα ομοιότητας συγκρίνονται ως προς την ποιότητα των παραγόμενων συστάδων με τη χρήση του δείκτη Rand, και συνεπώς εξάγονται συμπεράσματα για την αποτελεσματικότητα των μέτρων ύπο διαφορετικές συνθήκες.

CC BY: Attribution alone 4.0
https://creativecommons.org/licenses/by/4.0/
https://doi.org/10.26219/heal.aueb.5832

CC BY: Attribution alone 4.0
https://creativecommons.org/licenses/by/4.0/
https://doi.org/10.26219/heal.aueb.5832

TABLE OF CONTENTS

1 Introduction 1
1.1 Motivation 1
1.2 Objective 2
1.3 Organization 3
2 Background 5
2.1 Knowledge Discovery in Databases 5
2.2 Cluster Analysis Definition 7
2.3 Typical Clustering Steps 8
2.4 Clustering Methods 10
2.4.1 Partitioning Based Clustering Methods 10
2.4.2 Hierarchical Based Clustering Methods 10
2.4.3 Grid Based Clustering Methods 11
2.4.4 Model Based Clustering Methods 12
2.5 Similarity, Dissimilarity and Distance 12
2.6 Cluster Validation 14
2.7 Clustering Mixed Data 15
2.7.1 Methods for Clustering Mixed Data 15
2.7.2 Previous Work 16
2.7.3 Mixed Data Cluster Validation 18
2.8 Summary 21
3 Similarity Measures for Clustering Mixed Data 23
3.1 Gower's Similarity Coefficient and its Modifications 23
3.1.1 Gower's General Dissimilarity Coefficient 23
3.1.2 Inter-Quartile Gower's Measure Modification - D'Orazio 26

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

CC BY: Attribution alone 4.0
https://creativecommons.org/licenses/by/4.0/
https://doi.org/10.26219/heal.aueb.5832

3.1.3 Modified Kernel Density Estimation Distance - D'Orazio 27
3.1.4 Modified k-nn Distance - D'Orazio 27
3.1.5 Modification Based on 'Conditional Distance' - D'Orazio 28
3.1.6 Inverse Occurrence Frequency – Šulc, Z., Matejka, M., & Procházka, J. 29
3.1.7 Lin Measure - Šulc, Z., Matejka, M., & Procházka, J. 30
3.1.8 Variable Entropy Measure - Šulc, Z., Matejka, M., & Procházka, J. 30
3.1.9 Wishart's Dissimilarity Measure 31
3.1.10 Podani's Dissimilarity Measure 31
3.2 Generalization of the Gower's Coefficient with Weights 32
3.3 Other Similarity Measures 34
3.3.1 Laflin's General Similarity Coefficient 35
3.3.2 Huang Distance Measure (K-prototypes) 36
3.3.3 Ahmad and Dey Distance Measure (K-means) 36
3.3.4 Harikumar & PV Distance Measure 38
3.3.5 Modha and Spangler (convex K-means) 39
3.3.6 Similarity Based Agglomerative Clustering - Li & Biswas 40
3.3.7 TreeClust Dissimilarity 41
4 Simulation Study 43
4.1 Simulation Setup 43
4.2 First Experiment – Dataset Variation 46
4.2.1 Dataset Characteristics 46
4.2.2 Results for Dataset Difficulty Variation 50
4.2.3 Results for Minimal Between-Cluster Distances Variations 52
4.2.4 Results for the Combinations of Numeric and Categorical Variables 53
4.2.5 Conclusions 59
4.3 Second Experiment – High Number of Variables 59
5 Conclusions 61

CC BY: Attribution alone 4.0
https://creativecommons.org/licenses/by/4.0/
https://doi.org/10.26219/heal.aueb.5832

6 References 64

CC BY: Attribution alone 4.0
https://creativecommons.org/licenses/by/4.0/
https://doi.org/10.26219/heal.aueb.5832

CC BY: Attribution alone 4.0
https://creativecommons.org/licenses/by/4.0/
https://doi.org/10.26219/heal.aueb.5832

LIST OF TABLES

Table 2.1 Overview of linkage methods 11
Table 2.2 Dataset consisting of numeric, categorical and binary attributes 16
Table 3.1 Calculation of the Gower's Similarity by type of variable 25
Table 4.1 Measure symbolisms 45
Table 4.2 Dataset characteristics and combinations 48
Table 4.3 Descriptive statistics for example dataset. 49
Table 4.4 Average value of Rand index by dataset difficulty 51
Table 4.5 Variance of Rand index by dataset difficulty 52
Table 4.6 Average values of Rand index by minimal cluster distances. 53
Table 4.7 Variance of Rand index by minimal cluster distances. 53
Table 4.8 Average value of Rand index for four variables broken down by number of categorical variables. 54
Table 4.9 Variance of Rand index for four variables broken down by number of categorical variables. 55
Table 4.10. Average value of Rand index for six variables broken down by number of categorical variables. 55
Table 4.11 Variance of Rand index for six variables broken down by number of categorical variables. 56
Table 4.12 Average value of Rand index for eight variables broken down by number of categorical variables. 57
Table 4.13 Variance of Rand index for eight variables broken down by number of categorical variables. 58
Table 4.14 Average values for Rand index – Experiment 1. 59
Table 4.15 Average values for Rand index – Experiment 2 60
Table 4.16 Variance of Rand indices – Experiment 2 60

CC BY: Attribution alone 4.0
https://creativecommons.org/licenses/by/4.0/
https://doi.org/10.26219/heal.aueb.5832

CC BY: Attribution alone 4.0
https://creativecommons.org/licenses/by/4.0/
https://doi.org/10.26219/heal.aueb.5832

LIST OF FIGURES

Figure 2.1 Knowledge Discovery in Databases 5
Figure 2.2 Schematic diagram of a supervised and unsupervised learning algorithm 6
Figure 2.3 Clustering example: (a) input data and (b) the clusters 8
Figure 2.4 Sequential procedure of a cluster analysis process 9
Figure 4.1 Box plot for example dataset 49
Figure 4.2 PCA plot of example dataset 50

UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ
ATHENS

CC BY: Attribution alone 4.0
https://creativecommons.org/licenses/by/4.0/
https://doi.org/10.26219/heal.aueb.5832

CC BY: Attribution alone 4.0
https://creativecommons.org/licenses/by/4.0/
https://doi.org/10.26219/heal.aueb.5832

CHAPTER 1

Introduction

1.1 Motivation

Two main categories for data analysis are the supervised learning methods, in which the data has labels, and the unsupervised learning methods, in which no labelled training data exists. Cluster Analysis is an important unsupervised learning technique and is considered to be one of the fundamental techniques in data mining. It is described as the process of discovering representative groupings in a dataset. These groups are referred to as clusters, and each cluster represents a subset of the items that have been grouped in such a way that each item is linked to a single cluster. Clustering has been widely used in a variety of disciplines, including biology, psychology, statistics, psychiatry (Jain & Dubes,1988), healthcare systems (Mateo et al., 2008), finance (Liao et al., 2008) and customer relations management (Jing et al., 2007). A clustering method is considered efficient when objects within the same cluster are similar to each other but dissimilar to those in other groupings. For this reason, but primarily because a large category of clustering methods, that include the hierarchical clustering techniques, are based on similarity measures, a suitable measure of similarity or dissimilarity is needed. This way a partition structure can be identified in the form of natural groups.

For many years, the literature concentrated on clustering algorithms that could only handle numerical or categorical data. A numerical variable is typically measured in some number or value that corresponds to the attribute's quantity, while there is a natural ordering between distinct attribute values, and therefore they can be more or less similar to each other. Categorical variables, on the other hand, take on a value that represents one of several potential categories. The variables can take the form of text or numbers, but there is no natural order to the values in general. Categorical attributes can obtain an order which is encoded in to the values and in this case they are called ordinal variables.

When objects are characterised as numeric, there are several similarity/dissimilarity measures to choose from. The most widely used metric for this case is possibly the Euclidean distance, which is a subtype of the Minkowski distance (also referred to as the p-norm distance). In addition, in data mining applications, data are frequently characterised by categorical

CC BY: Attribution alone 4.0
https://creativecommons.org/licenses/by/4.0/
https://doi.org/10.26219/heal.aueb.5832

variables that accept values from a (usually) finite set of unordered nominal values. The simplest similarity/dissimilarity metric for categorical data is overlap (Kasif et al., 1998). For this, the proximity between two categorical objects is proportional to the number of attributes in which they match. Other metrics produced from overlap, such as the Jaccard coefficient, have been incorporated in many clustering algorithms (Huang, 1998).

Modern, real-world datasets though are often quite large and noisy, and usually contain data with mixed type attributes (i.e. numerical and categorical), and subsequently the main obstacle to clustering mixed data is determining how to unify the distance representation schemes for numeric and categorical data. It is difficult for applying traditional clustering algorithm directly into these kinds of data. Converting one kind of characteristic to another and then applying typical single-type clustering techniques may result in considerable information loss. As a result, it is preferable to overcome this problem by developing a unified similarity measure for categorical and numerical features, so that the metric gap between numerical and categorical data can be eliminated. Subsequently, a general clustering algorithm which is applicable to numerical and categorical data can be presented based on this unified metric. Many approaches for cluster formation in numeric environments have been suggested, such as agglomerative hierarchical clustering (Day & Edelsbrunner, 1984) and the classical k-means (Kaufman & Rousseew, 1990; Jing et al., 2007). Publications have also addressed the problem of clustering categorical data. Several researches have been published in recent decades that attempt to provide an uniform similarity metric for category and numerical variables.

There exist a few metrics that have been designed specifically for the purpose of clustering mixed data. One of the most popular measures is Gower's distance metric (Gower, 1971). Gower's coefficient will be discussed further in Chapter 3.

1.2 Objective

In this thesis we aim to explore several modifications for Gower similarity coefficient and various alterative unified measures of similarities for mixed-type data and study their performance. A comparison study was conducted including all available methods based on real and simulated data sets. The comparison of all similarity measures is performed by using agglomerative hierarchical clustering with a complete linkage method. The produced clusters are evaluated by the Rand Index.

CC BY: Attribution alone 4.0
https://creativecommons.org/licenses/by/4.0/
https://doi.org/10.26219/heal.aueb.5832

1.3 Organization

The rest of the thesis is organized as follows. In the next chapter, the background of Cluster Analysis is presented, while key features are described for different types of attributes. In Chapter 3 all the similarity measures evaluated in this thesis are presented. Chapter 4 contains the experiments conducted on simulated data. Finally, Chapter 5 contains a summary of the results and a discussion.

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

CC BY: Attribution alone 4.0
https://creativecommons.org/licenses/by/4.0/
https://doi.org/10.26219/heal.aueb.5832

CC BY: Attribution alone 4.0
https://creativecommons.org/licenses/by/4.0/
https://doi.org/10.26219/heal.aueb.5832

CHAPTER 2

Background

2.1 Knowledge Discovery in Databases

The world today is full of data, and everyday people face different types of data from a series of sources. Data supply a basis for further analysis, reasoning, decisions, and ultimately, for the understanding of all kinds of objects and phenomena (Xu & Wunsch, 2008, pp. 1). Over the last few years, data availability and usage have significantly increased as computer technology, which includes mass storage use, has become widely available and inexpensive (Brea,2013). The demand for the process and exploration of data and information has also escalated, leading to the necessity for new analytical approaches. The nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns or relationships within a dataset in order to make important decisions is known as Knowledge Discovery in Databases (KDD) (Fayyad, Piatetsky-Shapiro, & Smyth, 1996).

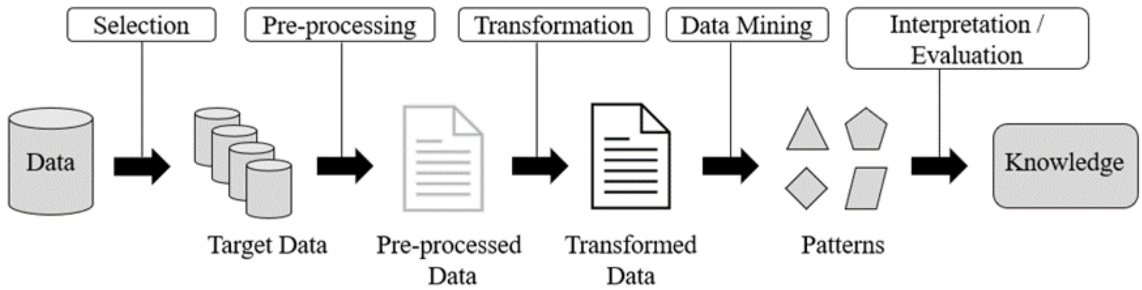

Figure 2.1 Knowledge Discovery in Databases (Fayyad, Piatetsky-Shapiro, & Smyth, 1996, pp. 84)

The steps of the knowledge discovery in databases processes are depicted in *Figure 2.1*. It is shown that from the data, specific characteristics are chosen so that the targeted data is processed by using the appropriate methods. The results from these methods create the final data patterns, that are interpreted/evaluated and sequentially formed into knowledge. Data analysis is one of the essential parts of Knowledge Discovery in Databases as different approaches, such as exploratory, statistical, predictive, etc., analyses data from different aspects and therefore the information found is visualized and interpreted in different forms (Hastie et al, 2009).

Traditionally, the two main methods that have been proposed in the knowledge discovery field are the supervised paradigm, and the unsupervised paradigm (Zhu & Goldberg, 2009).

UNIVERSITY OF ECONOMICS & BUSINESS · ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ · ATHENS

CC BY: Attribution alone 4.0
https://creativecommons.org/licenses/by/4.0/
https://doi.org/10.26219/heal.aueb.5832

The main difference between these two main methods is the existence of labels in the training data subset (Berry et al., 2019).

Supervision is based on the idea of a guide to the learning process. This occurs as supervised learning algorithms are based on the patterns learned by the existing data or historical datasets to predict the estimated output. Datasets are labelled as dependent and independent while they are used to solve problems involving classification and regression. Classification problems require finding a pattern to group the input dataset into distinct groups based on a specific pattern. The existing initial data labelling is used as a guide (training set) for the characteristics of every group so that every new data point is able to be submitted to one of these groups. Thus, classifications is based on the examination of the characteristics of a dataset, and based on these characteristics it is divided into a predetermined number of groups while the new data that is to be categorised is submitted into one of these predefined groups. Some techniques developed for this category are neural networks, decision trees etc. (Gan et al., 2020). Alternately to the classification techniques, regression problems predict the outcomes based on the relationship pattern identified between the input and output datasets (Breiman et al, 1984).

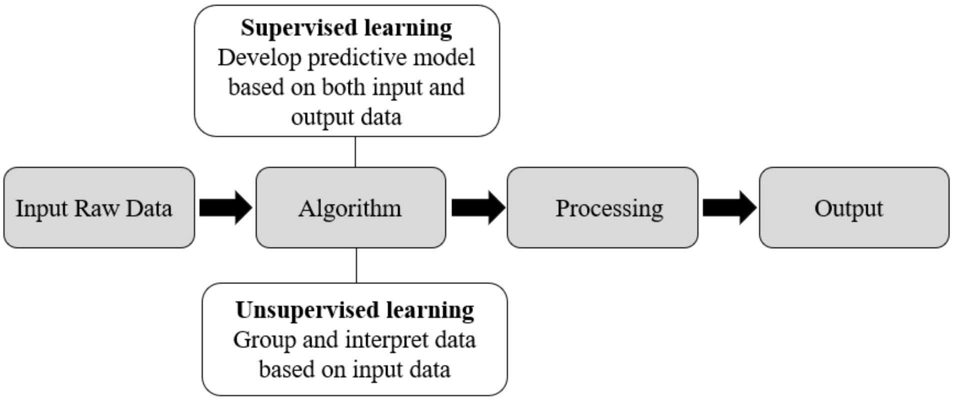

Figure 2.2 Schematic diagram of a supervised and unsupervised learning algorithm

On the other hand, this external guidance is unavailable in the case of unsupervised learning; thus the process of building a model from the data is harder than in the supervised paradigm (Cord & Cunningham, 2007, pp. 51-52). The collection of labelled data can be difficult, while it requires intensive human effort and can be highly time consuming, resulting in that this type of data is not broadly available (Dinler & Tural, 2016; Duda et al., 2001). *Figure 2.2* depicts a schematic diagram of a supervised and unsupervised learning algorithm as explained.

CC BY: Attribution alone 4.0
https://creativecommons.org/licenses/by/4.0/
https://doi.org/10.26219/heal.aueb.5832

One of the most popular methodologies in the case of unsupervised learning is clustering, which focuses on finding natural structures in the data that are not labelled (Dinler &Tural, 2016). Clustering has a long and rich history that stretches over a variety of scientific subjects such as biology, psychology, statistics, psychiatry, engineering, geology, geography, marketing and computer science (Jain & Dubes,1988; Celebi, 2015). Additionally, it can be endorsed into the steps of other research topics and applications such as customer segmentation, data exploration, data summarization, dynamic trend detection, information retrieval, market basket analysis, medical diagnostics, multimedia data analysis, social network analysis, text mining, and web analysis (Berkhin, 2006).

Within this thesis we will be focusing on Cluster Analysis and therefore we will include a more thorough review of this technique in the next part. We will review the literature on cluster analysis and cluster validation within this chapter. The first part focuses on the basic understanding of clustering methods. Then clustering with mixed-type data is introduced, and the appropriate methods are shortly presented.

2.2 Cluster Analysis Definition

A cluster in most definitions is described in terms of internal homogeneity and external separation (Gordon,1999). Specifically, cluster analysis is the task of dividing data objects into a number of groups based on their similarity, such that in each group the data objects are more similar to each other than the data objects from other groups. In other words, with a set of data objects, the primary problem of clustering is to partition it into a set of groups that are as similar as possible (Aggarwal, 2018). The groups that have similar data objects are referred to as clusters.

This is a subjective method as sets consisted by the same data may be clustered differently for different practices, resulting in the necessity of altering each stage of the method for an optimized performance. The key characteristic of the clustering concept is that in a data set, there are no unique unbiased perfect clusters. Clustering necessitates that the researchers define the type of clusters they aim for, which is determined by the subject matter and the goal of the clustering. Choosing an appropriate clustering method entails corresponding the data analytic features of the resulting clustering with the researcher's topic requirements. Most often the literature that introduces clustering methods is based on assumptions about what the true clusters are, which are rarely explained clearly. These assumptions are the following:

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

CC BY: Attribution alone 4.0
https://creativecommons.org/licenses/by/4.0/ https://doi.org/10.26219/heal.aueb.5832

1. Given a data set, there is a 'natural human conception' of what the true clusters are.
2. The data can be assumed to have been created by a statistical model in which different clusters are modelled by different distributions from the same family, or in which clusters are related to density modes.

In most circumstances, these assumptions are insufficient (Hennig & Liao, 2013).

In *Figure 2.3* *(a)* a sample dataset is shown in a 2-dimentional space, while *(b)* shows the clusters circled in the case that there is a total of three clusters in the dataset. The algorithm identified clusters for the objects that are very close to each other. In practice though, the datasets are not as simple as the one shown below and the data points are not always clearly separated between each other. As a result, the task of clustering is often more complex and challenging.

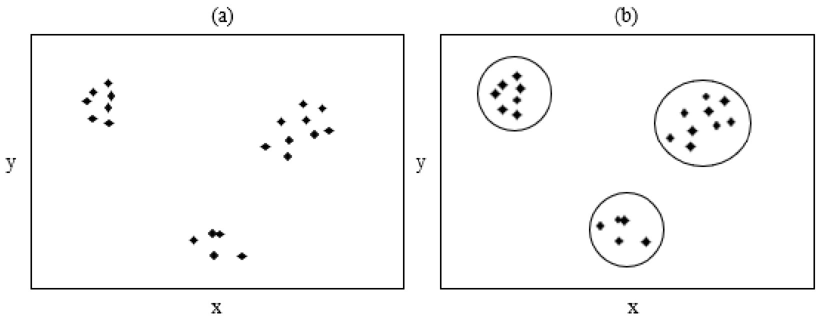

Figure 2.3 Clustering example: (a) input data and (b) the clusters

In the following section we discuss the fundamental steps for a typical clustering process.

2.3 Typical Clustering Steps

The clustering analysis methods usually follow a series of sequential steps (Heinz et al., 2003). According to Jain & Dubes (1988), the five main steps that most clustering algorithms use are: A) data representation B) definition of a proximity measure C) selection of a clustering algorithm D) data abstraction, and E) cluster validation. These steps are briefly described below and shown in *Figure 2.4*.

A) The representation of the data, which optionally includes feature extraction and/or selection. The representation of the data involves the number of clusters, the amount of available data and the number, type and scale of the available variables. Datasets often contain a large number of attributes to represent the data objects, but not all of these are always useful for the process. It is usually seen that some of the features are irrelevant to the cluster analysis. As mentioned by Witten et al. (2005), adding such attributes to the cluster

CC BY: Attribution alone 4.0
https://creativecommons.org/licenses/by/4.0/ https://doi.org/10.26219/heal.aueb.5832

analysis process undermines the performance of the technique used. Therefore, this process involves the recognition of the most useful subset of variables from the initial dataset variables (feature selection). Feature extraction involves the use of one or more variations of the input variables in order to create new variables. These methods can be used separately or simultaneously to create a total of variables that are suitable for use within the clustering method.

B) The selection of a proximity measure for the similarity between objects is directly involved with the creation of the clusters and therefore a definition of this measure should be defined. The proximity is usually measured by a distance function defined on pairs of data points. This is described in further detail in the next section.

C) The next step includes the clustering algorithm. Several algorithms partition the data into a predefined number of groups, whereas other algorithms output a nested series of clusters (Jain et al., 1999). Some of the algorithms are suitable for large datasets, whereas other methods deal with outliers better. As will be described in the next section, clustering can be divided into five main categories: Partitioning-based, Hierarchical-based, Density-based, Grid-based and Model-based (Vinh et al., 2009).

D) Data abstraction is a process of extracting a simple and compact representation of a dataset. The term simplicity concerns the ability of the produced dataset to be edited successfully from a machine, or by a human.

E) In the case where statistical techniques of clustering are used, the validation of the results takes place with careful use of statistical methods and hypotheses. As mentioned above, different measures and algorithms may conclude to different results. Therefore, it is necessary to compare, evaluate, and measure the goodness of the cluster analysis methods in order to achieve the best results. There are many different evaluation methods to help with this, that are also mentioned in this chapter.

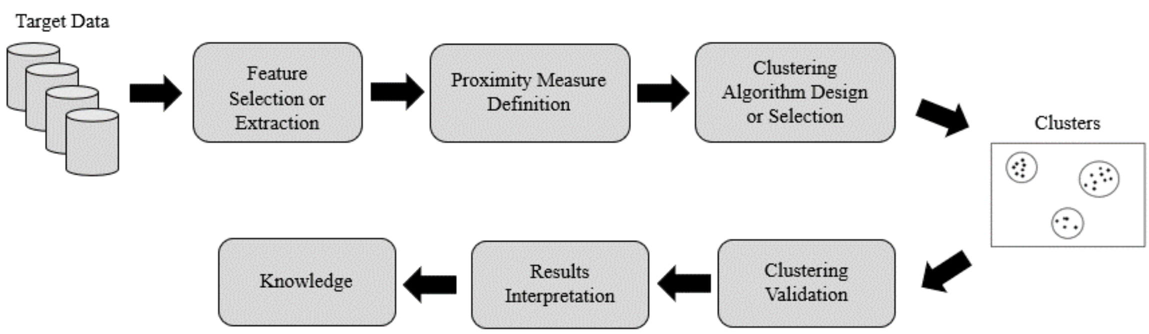

Figure 2.4 Sequential procedure of a cluster analysis process (Jain & Dubes, 1988)

CC BY: Attribution alone 4.0
https://creativecommons.org/licenses/by/4.0/ https://doi.org/10.26219/heal.aueb.5832

2.4 Clustering Methods

Clustering algorithms are broadly categorized as follows: Partitioning-based, Hierarchical-based, Density-based, Grid-based and Model-based (Vinh et al., 2009). These are introduced in this section.

2.4.1 Partitioning Based Clustering Methods

The partitioning based clustering methods, create partitions, the clusters, from a given set of data objects. Firstly, the data objects are assigned to the clusters, depending on some proximity measure, and then an iterative relocation process is used to improve the partitioning by moving the data objects from one cluster to another (Reynolds et al, 2006). It is essential in partitioning based methods that each cluster must contain at least one data object, and that each data object must belong to exactly one cluster (Pratima & Nimmakanti, 2013; Batra, 2013). The most well-known and commonly used partitioning methods are k-means, k-medoids, and their variations. Both of these methods imply that a centre point can represent a cluster. For k-means a notion of a centroid is used, which is the mean or median of all data objects in a partition. For k-medoid or PAM (Partition around medoids), the notion of a medoid is used, which is the most representative point of a cluster (Batra, 2013).

2.4.2 Hierarchical Based Clustering Methods

In contradiction with partitioning-based clustering which generates a flat partition of the data, hierarchical-based clustering is a method that attempts to build a hierarchy of clusters. Data objects are grouped into a tree of clusters, which is known as a dendrogram and shows the data partitioning at each stage. The tree root, which is a single cluster, consists of all the data objects, whilst more than one cluster is represented by the tree leaves. When the tree is terminated at a specific level, the data objects are clustered into disjointed groups (Pham & Afify, 2006).

This method can be broadly classified into divisive hierarchical clustering and agglomerative hierarchical clustering (Jain & Dubes, 1988). The divisive type is a top-down process, starting with a single cluster enclosing the total of the data points. This first cluster is then divided consecutively into two smaller groups until the termination criterion is established. The agglomerative type of hierarchical clustering adopts a bottom-up process, where each data point is initialised as a cluster, and works as follows. In the first stage, each object in a dataset is treated as a separate cluster. Based on a proximity matrix which includes

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

CC BY: Attribution alone 4.0
https://creativecommons.org/licenses/by/4.0/
https://doi.org/10.26219/heal.aueb.5832

dissimilarities for all pairs of clusters, a pair of objects with the lowest dissimilarity is merged into a new cluster. Then, a new proximity matrix between all clusters is calculated, and the algorithm repeats until there is only one cluster left, but can be cut off at any stage. A hierarchical structure is produced by using a certain between-cluster distance (Šulc & Řezanková, 2019). Merging of two clusters is accomplished at every repetition until the termination criterion is satisfied (Rafsanjani et al., 2012). The agglomerative type has an additional strength compared to the divisive type, as it allows analysis of the developed cluster hierarchy to establish the ideal number of clusters, as it is not obligatory that the final number of clusters are pre-defined (Gan et al., 2007).

The selection of a linking method, which establishes guidelines for defining between-cluster distances, is a crucial stage in Hierarchical Cluster Analysis. Some frequently used linking methods are only applicable to quantitative data such as Ward's method, which is applied only to the squared Euclidean distance, or the centroid method, which also can't be used on categorical data. The most commonly used linkage methods are the single, complete, and average linkage methods. The single-link (Sneath and Sokal, 1973) technique compares two clusters by calculating the highest similarity between the objects in each cluster. The complete-link technique (King, 1967), on the other hand, employs the lowest similarity between two objects to assess the same similarity. The average linkage method uses the average pairwise similarity between objects in two different clusters. It is a robust approach that is regarded as a middle ground between the single and complete linkage methods. An overview of these linkage methods is shown in Table 2.1, where n_g and n_h are the number of objects in clusters C_g and C_h (Šulc & Řezanková, 2019).

Linkage Method	$D(C_g, C_h)$
Single	$\min_{\mathbf{x}_i \in C_g, \mathbf{x}_j \in C_h} D(\mathbf{x}_i, \mathbf{x}_j)$
Complete	$\max_{\mathbf{x}_i \in C_g, \mathbf{x}_j \in C_h} D(\mathbf{x}_i, \mathbf{x}_j)$
Average	$\frac{1}{n_g n_h} \sum_{\mathbf{x}_i \in C_g \mathbf{x}_j \in C_h} D(\mathbf{x}_i, \mathbf{x}_j)$

Table 2.1 Overview of linkage methods

2.4.3 Grid Based Clustering Methods

Grid-based clustering quantize the object space into a finite number of cells that forms a grid structure (Jain et al., 1999). It forms this grid structure by dividing the data space into cells

CC BY: Attribution alone 4.0
https://creativecommons.org/licenses/by/4.0/
https://doi.org/10.26219/heal.aueb.5832

and then performs clustering on each grid cell instead of the database itself. The difficulty is based on the number of occupied grid cells and does not depend on the number of data points (Han, 2005). Since the size of the grid is usually much less than the number of the data objects, this method can have highly improved processing time. This method however does not perform well on highly irregular data.

2.4.4 Model Based Clustering Methods

Model-based clustering is one of the most frequently used clustering methods and is based on the assumption that data come from a population consisting of several subpopulations. Specifically, objects within a class hold a common distribution in their characteristics, and objects from different classes follow different distributions. This indicates that the whole population will follow a mixture of distributions, where the reason for clustering would be to take such a mixture and analyse it into simple components and estimate the probabilities that each observation belongs to each cluster (Zhang & Di, 2020). Each subpopulation is modelled separately and the total population is modelled as a mixture or weighted sum of these subpopulations, using finite mixture models (Raftery & Dean, 2006).

The choice the Clustering method, should be guided by each specific clustering problem, the type of data, the variables that characterise the data, and the similarity measure.

2.5 Similarity, Dissimilarity and Distance

As previously mentioned, two of the essential steps involved in Clustering analyses, that can be executed independently or concurrently are: the calculation of the similarity, or the distance between features or objects, and the application of an algorithm to expose latent clusters. Apart from some important exceptions such as mixture models (Hunt & Jorgensen, 2011) and modal clustering (Li et al., 2007), cluster analysis frequently depends on distance measures to quantify the similarity/dissimilarity between points in the feature space. Precisely, many clustering processes strive to minimise the distance between units in the same cluster while maximising the distance between units of different clusters (Foss et al., 2019).

Generally, we can think of a similarity coefficient as a way to calculate resemblance between two objects based on two sets of information leading to each individual with a number of variables for each set (Gower, 1971). As Everitt (1980) says, "the majority of clustering techniques begin with the calculation of a matrix of similarities and distances between

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

CC BY: Attribution alone 4.0
https://creativecommons.org/licenses/by/4.0/
https://doi.org/10.26219/heal.aueb.5832

entities, and therefore careful consideration is needed of the possible ways of defining these quantities".

These measures differ depending on the type of data included in the given dataset, resulting in a wide variety of different measuring techniques. For example, proximity measures that can be used successfully for numeric variables may not be appropriate for nominal data, due to the different representation of the feature values between the these data types.

As mentioned, the similarity is a numerical measure that represents the similarity between two objects. This measure usually returns a non-negative value that is between 0 and 1. In some situations, it may also range between -1 and +1 (Pearson Coefficient, Angular Separation, etc.). When the measure takes the value of zero, it means that there isn't any similarity between the objects, meaning that the data objects are very different from each other. In the contrary, when the measure take a value of one, the objects show complete similarity, implying that the objects are identical.

The dissimilarity measure is also a numerical measure, that shows the difference between two objects (Han et Kamber, 2006; Webb, 2022). This dissimilarity measure is reversely tied to the similarity measure, as when two data objects are very similar it will have a lower value, while when they are very different from each other it will have a higher value. As with the similarity measure, it can take value that fall between 0 and 1, while in some cases it can take values between -1 and +1.

Another term that can be used for these proximity measures, is the distance measure, which calculates the distance between two data objects in a multi-dimensional space (Meila et Shi, 2000). The distance measure takes values between 0 and ∞, while in some cases the distance measure and dissimilarity measure are considered as the same value. The properties that the distance measure follows are (Han et Kamber, 2006; Larose, 2004):

1. $d(x, y) = d(y, x)$ for all points x and y. For example, the distance from point x to point y is same as the distance from point y to point x.
2. $d(x, y) = 0$, if $x = y$. The distance will be 0 when the coordinates are the same.
3. $d(x, y) \geq 0$, for all points x and y.
4. $d(x, y) \leq d(x, z) + d(z, y)$, for all points x and y.

For a continuous dataset, the Minkowski Distance is a general method that is used to calculate the distance between two points. Specifically, the Manhattan Distance (Minkowski Distance of order 1) and the Euclidean (Minkowski Distance of order 2) are the most popular distance measures when dealing with continuous data. These distances are not data driven, as

OF ECONOMICS & BUSINESS ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ ATHENS UNIVERSITY ΟΙΚΟΝΟΜΙΚΟ

CC BY: Attribution alone 4.0
https://creativecommons.org/licenses/by/4.0/
https://doi.org/10.26219/heal.aueb.5832

the measures are independent of the data set of which the points came from. Other dissimilarity measures are cosine similarity measure, Pearson correlation, Mahalanobis distance. Cosine similarity measures the angle between two vectors regardless of vector magnitudes and Pearson correlation is often used in clustering gene expression data, although it is sensitive to outliers (Xu and Wunsch, 2005). Mahalanobis distance is scale-invariant, and takes into account variable correlations (Wang et al., 2021).

For categorical variables, simple matching dissimilarity is frequently used $d(\mathbf{x}_i, \mathbf{x}_{i'}) = \sum_{j=1}^{p} \delta(x_{ij}, x_{i'j})$, where p is the total number of variables and $\delta(x_{ij}, x_{i'j}) = I(x_{ij} \neq x_{i'j})$ showing whether variable j is the same for individuals i and i' (Wang et al., 2021).

The choice of distance measures is very important, as it has a strong influence on the clustering results. Depending on the type of the data and the researcher questions, other dissimilarity measures might be preferred.

2.6 Cluster Validation

As clustering algorithms result in uncovering partitions in the data, clustering validation methods are used for evaluating the found clusters in a quantitative and objective way (Jain & Dubes, 1988). Assessing the quality of a clustering solution is a necessary but difficult problem, as clustering is a subjective method. The main components for the validation is (1) determining whether there is reasoned/non-random structure in the data, (2) defining the number of clusters, (3) checking how well the clustering results fit the data when the data is the only information available and (4) evaluating how well a clustering solution agrees with partitions based on other data sources (Tan et al., 2005).

There are several evaluation and validation methods that can be used to achieve the above. Jain et al. (1998) suggest that these cluster validation measures are categorized into three groups, external, internal and relative measures. External validation methods use previous knowledge about the data, and therefore prior information of the data is obligatory, while internal validation methods rely only on the information derived from the data. The relative measures perform the evaluation by comparing the results from different cluster analysis methods with one another.

UNIVERSITY OF ECONOMICS & BUSINESS ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ ATHENS

CC BY: Attribution alone 4.0
https://creativecommons.org/licenses/by/4.0/
https://doi.org/10.26219/heal.aueb.5832

2.7 Clustering Mixed Data

Most clustering methods deal with either numerical data, represented by continuous values, or categorical data, which are a discrete type of data and can only have a finite number of values. As can be assumed, numerical data is common in datasets, but categorical data is also frequently found in real-life studies, such as name, gender, age group, educational level, and blood type (Dinh et al., 2021).

Thus, even though data that contains the same type of variables would be ideal, real-life data are often mixed type, meaning that in the dataset the selected variables are consisted by several different types of data. These sorts of datasets are already present in many disciplines, and with the appearance of big data, the availability of data sets comprised by heterogeneous sources and data types will keep on increasing (Fan et al., 2014). For example, medical data may consist of categorical variables, such as gender, job, education, marital status, and smoking or non-smoking, as well as numerical values such as age, height, and weight. Consequently, clustering mixed-type data is an important topic in cluster analysis, with numerous attempts made over the previous two decades to identify the proper approach for its solution.

2.7.1 Methods for Clustering Mixed Data

A variety of methods have been suggested for clustering mixed-type data, creating three main categories for solving this problem. The first approach is to analyse variables of each scale separately and to synthesize the results. This approach has a major drawback as the results from doing a separate clustering on each variable type may not agree (Kaufman & Rousseeuw, 1990). The second approach consists of methods that adapt already existing techniques for single-type data for their application to mixed-type data. This takes place with the transformation of the variables into one data type and may be done by converting all the categorical variables into continuous (Numerical Coding) or all the continuous variables into categorical (Discretisation) (Ahmad & Khan, 2019; Balaji & Lavanya, 2018). Therefore, if interval scale variables are discretised, clustering techniques for categorical data may be used, or if categorical variables are dummy coded, techniques for interval scale data may be used. Discretisation is a commonly used method, but issues may occur from its use as it involves a possibly significant loss of data resolution if unsuitable cut points are used (Kerber, 1992). Optimal intervals cannot be estimated easily, and boundary issues are created, as two close values near to the boundary may be separated into different ranges (Yang, 2012). The

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

CC BY: Attribution alone 4.0
https://creativecommons.org/licenses/by/4.0/
https://doi.org/10.26219/heal.aueb.5832

transformation of categorical values to a set of binary, numeric values suffers from an increase in dimensionality, as well as loss of semantics in categorical variables. For example, consider the sample of a dataset as shown in *Table 2.2*. In this dataset, Age is a numeric attribute, that has a natural ordering, as for example 35 > 18. Converting this numeric attribute into a categorical variable may assign 35 and 18 with the same categorical value. Therefore, such a transformation will lead to loss of important information. Marital Status is a categorical variable as there is no ordering between Single, Married and Divorced. Thus, it is difficult to transform categorical variables like Marital Status and Job to numeric values. Housing is a binary variable as its values are Yes and No, and therefore it also doesn't follow an ordering, as the numeric variables follow. In conclusion, each of the numeric, categorical and binary attributes have different characteristics and therefore should be approached differently. To compute the distance between two objects, for example between row 1 and row 2 in *Table 2.2*, a unified model of distance measure for such data needs to be used.

Id	Age	Job	Marital Status	Education	Housing
1	18	Management	Married	Tertiary	Yes
2	35	Technician	Divorced	Secondary	Yes
3	64	Technician	Married	Secondary	Yes
4	24	Technician	Single	Tertiary	No
5	82	Services	Single	Primary	Yes

Table 2.2 Dataset consisting of numeric, categorical and binary attributes

The third approach to cluster mixed type data, is based on a similarity measure which can incorporate both numeric and categorical variables. These measures can be used in a proximity matrix in hierarchical cluster analysis or in other multivariate methods. This approach does not include a transformation of the dataset, as the previous approach, making it very popular (Šulc et al., 2017). This is the method we will be focusing on within this thesis.

2.7.2 Previous Work

In 1971, J.C Gower introduced a coefficient to measure the similarity between two objects for mixed data, which is considered the best known dissimilarity measure of that type. The distance is always a number between 0 (identical) and 1 (maximally dissimilar), which makes this measure particularly appealing. Shortly, to calculate the Gower's dissimilarity measure between two items, each element is compared and a term is computed. When the Gower's coefficient handles nominal variables, it identifies the similarity between two categories of a

UNIVERSITY OF ECONOMICS & BUSINESS ATHENS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

CC BY: Attribution alone 4.0
https://creativecommons.org/licenses/by/4.0/ https://doi.org/10.26219/heal.aueb.5832

certain variable by using one for the categories that match and zero otherwise. For numeric variables, the Manhattan distance is used after scaling the variables to the unit range. The Gower distance is calculated as the average of the computed terms, which are the partial dissimilarities across individuals. In addition, the Gower's distance handles missing values and allows for adding a weighting scheme when averaging distances, in case a certain variable or group of variables is considered to be more important than others.

Although its appealing use and popularity, when using the Gower's similarity measure some issues may arise. The method used by the Gowers coefficient for handling nominal variables is thought of as a quite simplistic approach. In addition, in the absence of weights, as in the original measure, the categorical variables have a higher impact on the final distance than the numeric ones, because of the different nature of the variables. Appropriate weights can be selected based on the characteristics of the data under consideration. As mentioned by Gower (1971) though, "the decision on a rational set of weights is difficult".

New approaches to dissimilarity determination have been developed since the introduction of this coefficient. When Gowers coefficient handles nominal variables, it identifies the similarity between two categories of a certain variable by using one for the categories that match and zero otherwise. This is thought of as quite a simplistic approach, and therefore modifications of this coefficient have been suggested.

D'Orazio attempts to overcome the major limitations of Gower's distance by recommending certain changes in computing the distance on interval and ratio scaled variables. Simple modifications aim to lessen the influence of outliers on the scaled Manhattan distance, while other modifications, depending on kernel density estimation methods, aim to minimize the unbalanced contribution of the different types of variables (D'Orazio, 2021). Šulc et al. (2017) introduced three modification, which on contrary to the original Gower coefficient which only identifies if two categories match or not in the case of nominal variables, they offer different approaches to measuring the similarity between categories by using the Inverse Occurrence Frequency, the Lin Measure and the Variable Entropy Measure. Wishart (2003) and Podani (1999) suggested the use of two dissimilarity measures that are different to Gower's similarity coefficient when calculating numeric variables.

Apart from Gower's similarity measure and its modifications, there has been literature introducing other similarity measures. Laflin's coefficient takes into account the number of different variable types and results into each attribute having an equal contribution to the

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

CC BY: Attribution alone 4.0
https://creativecommons.org/licenses/by/4.0/
https://doi.org/10.26219/heal.aueb.5832

measure (Laflin, 1998). A k-means extension, suggested by Huang (1998), is the k-prototypes method, where the Euclidean measure is used for numeric data, and the simple matching coefficient for categorical data. The final distance measure is a linear combination of the two, where the influence of each component is determined by a parameter γ (Huang, 1998). Modha and Spangler (2003) found the optimal weight that minimizes the within-cluster weighted distance while maximizing the between-cluster weighted distance. They use the Euclidean distance for continuous variables with the Cosine distance for categorical variables (Modha & Spangler, 2003). Harikumar and PV (2015) propose a generalized distance function in the form of a triplet that consists of three different distance measures for numeric, categorical and binary data types. Ahmad and Dey (2007) established the distance between categorical attribute values based on an attribute's ability to separate data into homogeneous segments and calculated the weights for numeric attributes by using the co-occurrence relation. The total distance is calculated by adding the categorical and weighted numerical distances. Li and Biswas (2002) suggest the Similarity-Based Agglomerative Clustering (SBAC) algorithm for mixed data clustering. SBAC uses Goodall similarity measure (Goodall, 1966), which gives more weight to uncommon feature value matches, and applies group-average methods for the aggregation (Jain & Dubes, 1988). The TreeClust Dissimilarity fits a set of regression or classification trees with each variable acting in turn as the “response" variable, where two observations are similar if they tend to fall in the same leaves of classification or regression trees (Buttrey & Whitaker, 2016).

This thesis focuses on dissimilarity measures for mixed data in hierarchical cluster analysis (HCA). The aim of this thesis is to explore several extensions of the Gower’s similarity coefficient and alternative measures of similarity for mixed-type data and study their performance.

2.7.3 Mixed Data Cluster Validation

The majority of internal indices are based on the original data matrix X, which in the case of mixed-type data is not able to be computed as it requires numeric variables. There are some internal indices though that do not relay on the original data matrix X, but on the dissimilarity matrix. Some of these are the Dunn, Silhouette and McClain indices.

The Dunn index (Dunn, 1974) supposes that clusters in a dataset are compact and effectively separated by maximizing the inter-cluster distance while minimizing the intra-cluster distance (Yang, 2012). Calculation of the Dunn index requires the determination of the distance

UNIVERSITY OF ECONOMICS & BUSINESS ATHENS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

CC BY: Attribution alone 4.0
https://creativecommons.org/licenses/by/4.0/ https://doi.org/10.26219/heal.aueb.5832

between two clusters, which is given by the distance between their closest points. For the cluster solution with k clusters, it can be expressed by the formula

$$Dunn\ (k) = \min_{1 \le g < h \le k} \left(\frac{d(g,h)}{\max\limits_{1 \le m \le k} \text{diam}\ (m)} \right) \tag{2.1}$$

where $d(g,h)$ is a distance between the g-th and h-th clusters expressed by one of the between cluster distances, diam (m) is the maximal distance between two objects in the m-th cluster. The Dunn index takes values from zero to infinity. The highest value indicates the optimal cluster solution.

The silhouette index (Rousseeuw, 1986), also known as average silhouette width can be written as:

$$\text{silhouette}\ (k) = \frac{1}{n} \sum_{i=1}^{n} \frac{b(i) - a(i)}{max[a(i), b(i)]} \tag{2.2}$$

where n is the number of objects, $a(i)$ is the average dissimilarity of the i-th object to the other objects in the same cluster, $b(i)$ is the minimal average dissimilarity of the i-th object to any cluster not containing i. Its value ranges from -1 to 1. Silhouette coefficients close to 1 means the sample is far away from the neighbouring clusters and therefore the clusters are well separated. A value of 0 means that the sample is on or very close to the border between two neighbouring clusters, while negative values indicate that the samples could have potentially been assigned to the wrong cluster (Rousseeuw, 1987).

The McClain index (McClain and Rao, 1975) is defined as a ratio of the within-cluster and between-cluster distances

$$\text{McClain}\ (k) = \frac{S_w / n_w}{S_b / n_b} = \frac{S_w n_b}{S_b n_w} \tag{2.3}$$

where n_w is the number of pairs of objects in the same cluster in the dataset, n_b is the number of pairs of objects not belonging to the same cluster, S_w is the sum of the within cluster distances between n_w pairs of objects, and S_b is the sum of the between-cluster distances between n_b pairs of objects. A lower value of the index indicates for a better cluster solution.

For external validation purposes the Rand index is used (Rand, 1971). In the case of external validation, as previously mentioned, prior knowledge regarding the groupings of the data is necessary. The prior knowledge concerns the pre-defined groupings of the data, and can be

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

CC BY: Attribution alone 4.0
https://creativecommons.org/licenses/by/4.0/ https://doi.org/10.26219/heal.aueb.5832

used as a reference when determining if the objects in the produced cluster solution are correctly assigned.

The computation of this index involves the formula for counting the pairwise combinations which is, given N objects, the total number of pairs that can be made from these N objects and is calculated as $N(N-1)/2$, and four possible states may result. A true positive (TP) or true negative (TN) state happens when an objects is assigned to a cluster that agrees with the reference grouping, where true positive is the number of pairs of instances that are assigned to the same cluster in both reference and produced clusters, while true negative is the number of pairs of instances that are assigned to a different cluster in both reference and produced clusters. A false positive (FP) or false negative (FN) state, on the contrary, occurs when an object is assigned to a cluster that does not agree with the reference grouping. Similarly, false positive is the number of pairs of instances that are in the same cluster in the reference groupings but in different clusters in the produced clusters, while a false negative is the number of pairs of instances that are in different clusters in the reference clusters, but in the same clusters in the produced clusters. Then the formula is:

$$\text{Rand } = \frac{TP + TN}{TP + FP + TN + FN} \tag{2.4}$$

The Rand index represents a ratio of correctly assigned objects, both positively and negatively, out of the number of all possible pairs. It takes values from 0 to 1 and values that are nearer to 1 mean better accuracy of the cluster assignment.

The Rand Index has one disadvantage in that its expected value between two random clusters may be different. The Adjusted Rand Index would be used to compensate for this, as it accounts for chance when measuring the proportions (Milligan and Cooper, 1985). The Adjusted Rand Index is given by (Talburt, 2011):

$$\text{Adjusted Rand Index } = \frac{TP - \left(\frac{(FP + TP) \cdot (FN + TP)}{TP + FP + FN + TN}\right)}{\frac{(FP + FN + 2TP)}{2} - \left(\frac{(FP + TP) \cdot (FN + TP)}{TP + FP + FN + TN}\right)} \tag{2.5}$$

It is noted that although the Rand index is always positive, the Adjusted Rand Index takes values between -1 and 1, where 1 is the perfect score (Mollaian et al., 2022). The examined dissimilarity measures for this thesis will be compared and evaluated regarding their clustering abilities measured by the Rand index.

UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ
ATHENS

CC BY: Attribution alone 4.0
https://doi.org/10.26219/heal.aueb.5832
https://creativecommons.org/licenses/by/4.0/

2.8 Summary

In spite of the abundance of clustering techniques and algorithms, clustering mixed interval and categorical scale data remain a challenging problem. In this study we will examine the Gower coefficient for mixed data and its modifications. We will also evaluate some of the other similarity measures used for mixed data clustering.

CC BY: Attribution alone 4.0
https://creativecommons.org/licenses/by/4.0/
https://doi.org/10.26219/heal.aueb.5832

CC BY: Attribution alone 4.0
https://creativecommons.org/licenses/by/4.0/
https://doi.org/10.26219/heal.aueb.5832

CHAPTER 3

Similarity Measures for Clustering Mixed Data

In this section, the properties of several unified similarity measures for mixed data are reviewed. We begin with Gower's Similarity Coefficient, as it is one of the similarity coefficients that have been theorized and described extensively, and then proceed to several of its variations. Lastly, we include other similarity measures suggested from the literature.

3.1 Gower's Similarity Coefficient and its Modifications

3.1.1 Gower's General Dissimilarity Coefficient

As previously mentioned, a detrimental point for identifying clusters in a dataset with mixed type variables is finding a distance metric that is able to work with its different data types. Gower's proposal in 1971 is the most used way in defining this measure. If we consider two distinct individuals as i and j, each having p features, the base for the Gower's Similarity Coefficient is given by:

$$s_{G,ij} = \frac{\sum_{t=1}^{p} \delta_{ijt} d_{ijt}}{\sum_{t=1}^{p} \delta_{ijt}} \qquad (3.1)$$

where $s_{G,ij}$ is the similarity between i and j, d_{ijt} is a dissimilarity measure between the i-th and j-th objects by the t-th variable, p is the total number of variables, and δ_{ijt} takes the value of zero, if either the i-th and j-th objects by the t-th variable is missing, while otherwise it takes a value of one.

The Gower's distance was originally introduced as a similarity measure but it can also be defined as the compliment to one of the Gower's similarity coefficient:

$$d_{G,ij} = 1 - s_{G,ij} = \frac{\sum_{t=1}^{p} \delta_{ijt} d_{ijt}}{\sum_{t=1}^{p} \delta_{ijt}} \qquad (3.2)$$

This is a dissimilarity or distance measure (Kaufman and Rousseeuw, 1990) between the two data objects i and j, where t takes values in a set of attributes of total size p, $d_{ijt} = 1 - s_{ijt}$ is the distance computed on the t-th variable and s_{ijt} is the similarity between i and j and depends on the type of variable.

CC BY: Attribution alone 4.0
https://creativecommons.org/licenses/by/4.0/
https://doi.org/10.26219/heal.aueb.5832

This formula is applied to data in a dataset with p columns since individuals i and j can be translated into data records x_i or x_j from the dataset $\boldsymbol{X}$. For this reason we mention data records x_i and x_j from a dataset $\boldsymbol{X}$ having their values about a specific feature t being compared (both values represented by x_{it} and x_{jt} respectively) rather than individuals i and j.

The indicator $\delta_{G,ij} = 0$ *if,*

i) x_{it} or x_{jt} or both are missing (x is the attribute value) or
ii) $x_{it} = x_{jt} = 0$ and t is an Asymmetric Binary attribute. An Asymmetric Binary attribute, is that for which the two states are not equally important, such as the positive (1) and negative (0) outcomes of a disease test. Given two Asymmetric Binary attributes, the agreement of two 1s (a positive match) is then considered more significant than that of two 0s (a negative match) (Han & Pei, 2011).

Otherwise, $\delta_{G,ij} = 1$.

The value of d_{ijt}*:* In this formula the distance between object i and j for a variable t is calculated using different distance measures that have already been established for various attribute types. A few examples are shown below.

- Suppose t is Numeric, then the Manhattan distance is used which has the formula: $d_{ijt} = \frac{|x_{it}-x_{jt}|}{max(x_{t})-min(x_t)} = \frac{|x_{it}-x_{jt}|}{R_t}$, where R_t is the range of the t-th variable.
- Suppose t is Binary Symmetric or Categorical Nominal then the Hamming distance is used: if $x_{it} = x_{jt}$ then $d_{ijt} = 0$, otherwise $d_{ijt} = 1$.
- Suppose t is Ordinal: The first Gower's coefficient method does not define a specific way for categorical ordered variables. Kaufman and Rousseeuw (1990) suggested that the categories of the variable should be substituted with the corresponding position index in the natural ordering. First we calculate the rank r_{it} for object i, suggesting that the attribute t has M_t ordered states and $r_{it} \in 1, \dots M_t$. This takes place by replacing x_{it} with its corresponding rank. This is controlled by the importance of the attribute value. A value with the most weight will have the highest rank whereas an attribute with less weight will have a lower rank. When the ranking has taken place, value normalization is necessary so that each of them falls in the interval [0.0, 1.0], and is achieved with the following formula: $z_{it} = \frac{r_{ij}-1}{M_t-1}$. There, z_{it} is considered to be a

CC BY: Attribution alone 4.0 https://doi.org/10.26219/heal.aueb.5832
https://creativecommons.org/licenses/by/4.0/

numeric variable and a distance metric for the numeric variable is used to calculate the distance between the objects.

The following table is used to summarise these categories (D'Orazio, 2021).

Variable Type	s_{ijt}	δ_{ijt}	Note
Binary Symmetric	1 if $x_{it} = x_{jt}$ 0 if $x_{it} \neq x_{jt}$ 0 if x_{it} or x_{it} or both are missing	1 if both the variables are not missing 0 if x_{it} or x_{it} or both are missing	s_{ijt} corresponds to the simple matching coefficient
Binary Asymmetric	1 if $x_{it} = x_{jt} = 1$ 0 otherwise 0 if x_{it} or x_{it} or both are missing	1 if both the variables are not missing 0 if $x_{it} = x_{jt} = 0$ 0 if x_{it} or x_{it} or both are missing	s_{ijt}corresponds to the Jaccard index
Categorical Nominal (more than two categories)	1 if $x_{it} = x_{jt}$ 0 if $x_{it} \neq x_{jt}$ 0 if x_{it} or x_{jt} or both are missing	1 if both the variables are not missing 0 if x_{it} or x_{jt} or both are missing	s_{ijt} is the simple matching on the untransformed variable
Measured on an interval scale	$1 - \lvert x_{it} - x_{jt} \rvert / R_t$ 0 if x_{it} or x_{jt} or both are missing	1 if both the variables are not missing 0 if x_{it} or x_{jt} or both are missing	$R_t = max(x_t) - min(x_t)$ is the range of the k-th variable $1 - s_{ijt}$ is the Manhattan distance scaled by the range

Table 3.1 Calculation of the Gower's Similarity by type of variable (D'Orazio, 2021, pp. 5).

In general, the Gower's dissimilarity measure, can be thought of as an average of the dissimilarities measured on the available variables, where the dissimilarity calculated on each single variable is scaled to range from 0, for the most different units, to 1, for the most similar, and as a result the averaging will give an overall dissimilarity taking values between 0 and 1 $\left(0 \leq d_{G,ij} \leq 1\right)$.

Another characteristic of Gower's dissimilarity coefficient is that it allows for missing values. These values are not considered in the calculation of the overall dissimilarity measure. This implies that if a unit is consisted only by missing values then the dissimilarity with any other unit which is partly observed will be undefined. Consequently, a unit with all the values missing should be removed from the dataset before the measure is calculated.

Gower's coefficient also allows the selection of weights for each individual variable, and therefore modifying the significance of each of them. This takes place when it is believed that one variable or a set of variables is more important than others. Weighting may be achieved by either giving different weights to each variable by multiplying the numerator and denominator by the weight, or by giving a weight to a specific comparison when values produce a certain property while being compared, by transforming the weights into a function that depends on these values (Gower, 1971). These weights can be selected, but are often set to one, as in the Gower's coefficient. The selection of weights is discussed in greater detail in *Chapter 3.2*.

CC BY: Attribution alone 4.0
https://creativecommons.org/licenses/by/4.0/ https://doi.org/10.26219/heal.aueb.5832

The unweighted Gower's distance assigns the same weight to each chosen variable. This standard unweighted Gower's distance shows an unbalanced contribution of quantitative variables compared to categorical ones. The unbalanced contribution of the different type of variables is further exacerbated when the number of categorical variables increases. In addition, the distance of interval/ratio scale variables is 0 only when $x_{it} = x_{jt}$ and 1 when $|x_{it} - x_{jt}| = R_t$, which is rare in the present of outliers as the estimation of the range R_t is highly affected by them. As shown in experiment performed by Šulc et al (2016), the original Gower measure provided good results; however, with the increasing number of categorical variables, its clusters worsened substantially. Therefore, it is recommended by them that this measure can be used for mixed data sets with the a low number of categorical variables.

Gower's coefficient has many advantages as discussed, and therefore is the most popular similarity measure for mixed data. Due to the above disadvantages though, many modification of this measure have been suggested. Specifically, these methods try to account for the presence of outliers in the quantitative variables, for the skewness in the distribution and attempt to better balance the contribution of the different types of variables on the overall unweighted Gower's distance. Some of these modifications are described in the next sections.

3.1.2 Inter-Quartile Gower's Measure Modification - D'Orazio

A fist modification that is mentioned by D'Orazio (2021), is a change applied on the interval/ratio scaled variables. It is achieved by replacing the range R_t with the inter-quartile range as it uses the following formula.

$$d_{ijt} = \begin{cases} \frac{|x_{it} - x_{jt}|}{IQR_t}, \text{if } |x_{it} - x_{jt}| < IQR_t \\ 1, \text{otherwise} \end{cases} \tag{3.3}$$

where IQR_t is the inter-quartile range, P75% - P25%. This is a simple modification that tries to solve the impact of outliers in the interval or ratio-scaled variables. Instead of this, the inter-decile, could also be used as a replacement.

Although this modification allows the determination of outliers, it does not solve the dominance of categorical variables, compared to interval variables, on the overall distance. D'Orazio concludes, through the results obtained in the simulation studies that were performed, that in general, if in presence of outliers, scaling the Manhattan distance by the Inter-Quartile range

CC BY: Attribution alone 4.0
https://creativecommons.org/licenses/by/4.0/
https://doi.org/10.26219/heal.aueb.5832

tends to perform better than by scaling by the range; specifically in maintaining the true distribution.

3.1.3 Modified Kernel Density Estimation Distance - D'Orazio

One way of partly solving the unbalanced contribution of the different types of variables, is by discretizing the interval and ratio-scaled variables. There are many different discretization criterions that can be used (equal-width classes, equal-frequency classes, etc.). An issue when using fixed discretization is that the observations that belong to two side by side classes can wrongly be closer than units belonging to the same class but at opposite boundaries. Therefore, D'Orazio proposes a non-fixed discretization, which follows the kernel density estimation theory. This means that two units that fall within the same moving 'window' will be considered being of distance equal to 0, and two units that one or both do not fall in the same 'window' will be calculated using the original formula. This is:

$$d_{ijt}^{(kde)} = \begin{cases} 0, & \text{if } \left|x_{it} - x_{jt}\right| \leq h_t \\ \dfrac{\left|x_{it} - x_{jt}\right|}{g_t}, & \text{if } h_t < \left|x_{it} - x_{jt}\right| < g_t \\ 1, & \left|x_{it} - x_{jt}\right| \geq g_t \end{cases} \tag{3.4}$$

where g_t can be the range of the t variable as in the standard Gower's distance, or the inter-quartile, h_t is the window width, defined as the bandwidth in the kernel density estimation. In this case h_t is considered to be the following, as defined by Silverman in 1986.

$$h_t = c \frac{1}{n^{1/5}} min\left\{s_t, \frac{IQR_t}{1.34}\right\} \tag{3.5}$$

where s_t is the estimated standard deviation for the t-th variable. Silverman (1986, p.48) mentions that $c = 1.06$ is quite precise when working with skewed unimodal distributions whereas $c = 0.9$ has a better performance for skewed and moderately bimodal distributions.

Through the simulations performed in his studies, D'Orazio found that this modification, seems to perform slightly better for all the assessment criteria (apart form one) compared to the other modifications he proposes.

3.1.4 Modified k-nn Distance - D'Orazio

Another modification by using the kernel density estimator philosophy (D'Orazio, 2021), uses the k-nearest neighbour method (k-nn), instead of the window h_t as a criterion to define classes.

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

CC BY: Attribution alone 4.0
https://creativecommons.org/licenses/by/4.0/
https://doi.org/10.26219/heal.aueb.5832

$$d_{ijt}^{(knn)} = \begin{cases} 0, & \text{if } x_{jt} \text{ is one of the } k \text{ nearest neighbors of } x_{it} \\ \dfrac{|x_{it} - x_{jt}|}{g_t}, & x_{jt} \text{ not one of } k \text{ nearest neighbors of } x_{it} \; and |x_{it} - x_{jt}| < g_t \\ 1, & \text{if } |x_{it} - x_{jt}| \geq g_t \end{cases} \tag{3.6}$$

where $k = \sqrt{\text{n}}$. The k value indicates the count of the nearest neighbours while n is the number of objects in the training dataset. The choice of k is very critical, as a small value means that noise will have a higher influence on the result, whereas a large value will lead to a more computationally expensive process. D'Orazio suggests using the well-known rule of thumb in which k equals the square root of n, as consistency may be obtained this way.

D'Orazio states that the knn-based method is known to perform better than the kernel density in the tails of the distribution. Therefore, $d_{ijt}^{(knn)}$ may be preferable to $d_{ijt}^{(kde)}$ when measuring distance on continuous variables with a highly skewed distribution with a long tail. In addition, when the number of continuous variables increases, the modified distances based on the k-nn and kernel density estimation appear to perform slightly better in retaining the marginal distribution of the target variable, while they also perform well in detecting a 'true' value when there are missing values in the data.

3.1.5 Modification Based on 'Conditional Distance' - D'Orazio

Another way to deal with mixed-type data when using distance-based measures in general, is by calculating the conditional distances. This means that only the distances withing each group will be calculated and not the distances between different groups. Firstly, chosen categorical variables of the same value are used to form classes. Then the distance between the values in the same group is calculated by using the remaining quantitative variables.

D'Orazio (2021) suggests an approach in which firstly, the distance using the Gower's formula for the subset of p_{cat}, the total number of binary or categorical nominal variables within the dataset, is calculated. Then the distance for interval or ratio scaled variables is calculated by using the Gower formula (Manhattan distance that is scaled by the range or IQR) only between observations whose first calculated distance is smaller than $1/p_{cat}$. If this is not satisfied by any of the units, then this shall increase to $2/p_{cat}$ etc..

Therefore, the distance between two units is set equal to Gower's distance, as stated above, for units that satisfy the condition. The remaining distances are equal to 1 which is the maximum

CC BY: Attribution alone 4.0
https://creativecommons.org/licenses/by/4.0/
https://doi.org/10.26219/heal.aueb.5832

value. By using this method, the distance on the interval or ratio-scaled variables is calculated only between the observations that show a difference in one (or two) of the categorical variables.

The result of the 'Conditional Distance' method is that categorical variables have higher importance in the creation of the initial groups. This is thought of as a very efficient modification from computational purposes since it reduces the effort in calculation of overall distances. While comparing the modifications, D'Orazio found that the 'Conditional Distance' approach seemed to be more effective, in reducing the unbalanced contribution of the different types of variables in the overall distance when the continuous variables were characterised by a skewed distribution (D'Orazio, 2021). However, this proposal does not perform well in reproducing the "true" deleted values, when missing values are present.

With regards to the modifications proposed by D'Orazio (2021), the results obtained by his studies are very close for all mentioned and no clear outperformance is seen.

3.1.6 Inverse Occurrence Frequency – Šulc, Z., Matejka, M., & Procházka, J.

This modification of the Gower's coefficient is based on the Inverse Occurrence Frequency (IOF) measure and focuses on improving nominal data by including additional information compared to the original approach of Gower's coefficient (Šulc et al., 2016). OF (occurrence frequency) and IOF (inverse occurrence frequency) take into account the absolute frequencies of the included categories in order to accomplish a more precise similarity definition between two categories in the case of mismatches. This measure was initially introduced by Sparck-Jones in 1972, as a similarity measure in the field of information retrieval, to find the amount of documents that include a certain word.

This modification assigns higher weights to less frequent non-matches for nominal variables as it is computed with the following formula:

$$d_{ijt} = \begin{cases} 0 & \text{if } x_{it} = x_{jt} \\ 1 - \dfrac{1}{1 + \ln f(x_{it}) \cdot \ln f(x_{jt})} & \text{otherwise} \end{cases} \tag{3.7}$$

where $f(x_{it})$ is an absolute frequency of the value x_{it} in the t-th variable. This results to zero when the categories match, and from zero to the number $\left(1 - 1/(1 + \ln{(n/2)^2})\right)$ if they don't match. As the dataset size n increases, the upper border of this range converges to one.

In the experiments that were performed by Šulc et al., the clusters produced by the Inverse Occurrence Frequency modification seem to not perform well from the aspects of increasing

CC BY: Attribution alone 4.0
https://creativecommons.org/licenses/by/4.0/
https://doi.org/10.26219/heal.aueb.5832

number of categorical variables and classification, and therefore the method is not suggested for common use.

3.1.7 Lin Measure - Šulc, Z., Matejka, M., & Procházka, J.

Another Gower's modification mentioned by Šulc et al. (2016) is achieved by using the measure introduced by Lin in 1998. This measure, when calculating the distance for nominal variables, assigns higher weights to non-matches to less frequent categories with the formula:

$$d_{ijt} = \begin{cases} 0 & \text{if } x_{it} = x_{jt} \\ 1 - \dfrac{2 \cdot \ln\left(p(x_{it}) + p(x_{jt})\right)}{\ln\, p(x_{it}) + \ln\, p(x_{jt})} & \text{otherwise} \end{cases}, \tag{3.8}$$

where $p(x_{it})$ is a relative frequency of the value x_{it} in the t-th variable. The results of the measure is a value between zero and one. The value of one is resulted if there are only two categories in a variable, and its limit is close to zero if both the observed categories have very small relative frequencies (Šulc, 2016).

Šulc et al. mention that through the experiments performed, the Lin Measure modification proved to be the most stable with regards to the increasing number of categorical variables. Since the cluster solution with three categorical variables, the Lin Measure modification is shown to outperform the original Gower measure. It is also observed that its classification performed well. Therefore, this measure can be recommended for datasets with a high number of categorical variables.

3.1.8 Variable Entropy Measure - Šulc, Z., Matejka, M., & Procházka, J.

In 2015, Šulc and Rezanková introduced the Variable Entropy similarity measure, which is the base of a modification of Gower's coefficient introduced by Šulc et al. in 2016. For nominal variables, matches that have high variability, which is shown by the normalized entropy, will have higher values of dissimilarities. This happens as variables with higher variability are rarer than matches in the low-variability variables. This measure can be shown with the formula

$$d_{ijt} = \begin{cases} 1 + \dfrac{1}{\ln K} \displaystyle\sum_{u=1}^{K_t} p_u \ln\, p_u & \text{if } x_{it} = x_{jt} \\ 1 & \text{otherwise} \end{cases} \tag{3.9}$$

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

CC BY: Attribution alone 4.0
https://creativecommons.org/licenses/by/4.0/
https://doi.org/10.26219/heal.aueb.5832

where $u(u = 1, \dots, K_t)$ is a category in the t-th variable, and p_u is a relative frequency of the t-th variable. It can take values equal to or between zero and one. Values closer to zero suggest lower dissimilarity for the categories.

In the Šulc et al. experiments, it is resulted that the clustering performance of the Variable Entropy Measure modification is preferred compared to the other examined measures in all combinations of numeric and categorical variables. Only in situations, when all the variables are categorical, it is outperformed by the Lin Measure modification, but this does not involve mixed variables. Based on the performance of the original Gower measure, a rapid decrease of performance is seen with increasing numbers of categorical variables, which is not observed for the Variable Entropy Measure modification, where the decrease is much lower. Therefore, this dissimilarity measure can be recommended for general use.

3.1.9 Wishart's Dissimilarity Measure

Another measure was proposed by Wishart (2003), which was used for the modification of the Gower similarity in order to include it in a k-means algorithm. This measure is different to Gower's similarity coefficient when calculating numeric variables, as it substitutes the range with a variance weight of the numeric variable in the s_{ijt} (see Equation 3.1), while it also applies a squared distance component, resulting in the following distance:

$$d_{ij} = \sqrt{\sum_{t=1}^{p} \delta_{ijt} \left(\frac{x_{it} - x_{jt}}{w_{ijt}} \right)^2} \tag{3.10}$$

where $\delta_{ijt} = 0$ if objects i and j for variable t is invalid due to lack of data, otherwise $\delta_{ijt} = 1$. For binary and categorical variables $w_{ijt} = 1$ if $x_{it} = x_{jt}$ and 0 otherwise, and for continuous variables $w_{ijt} = \sigma_t$, when t is a numerical variable, σ_t is the standard deviation of the t-th variable.

The Wishart distance has similar properties to the Gower distance such that the same assumptions apply, missing values may be included and non-equal contribution of variables may be defined.

3.1.10 Podani's Dissimilarity Measure

Podani (1999) introduced another measure as an extension to Gower's coefficient, in order to be able to incorporate ordinal type variables, which is not applicable by Gower's similarity

CC BY: Attribution alone 4.0
https://creativecommons.org/licenses/by/4.0/
https://doi.org/10.26219/heal.aueb.5832

measure. While the Gower index treats the ordinal variables as quantitative variables rescaled for their range, Podani extends the Gower index to ordinal variables considering a distance-based on their rank. Specifically, this is achieved by incorporating differences in ranks of two objects within the same rank order. For ordinal variable all x_{it} are replaced by their ranks r_{it} determined over all objects, and then:

$$s_{ijt} = 1 - \frac{|r_{it} - r_{jt}|}{max\{r_k\} - min\{r_k\}} \tag{3.11}$$

And the distance measure is found with the following formula:

$$d_{ij} = \sqrt{\sum_{t=1}^{p} \delta_{ijt} \left(\frac{x_{it} - x_{jt}}{w_{ijt}} \right)^2} \tag{3.12}$$

where $\delta_{ijt} = 0$ if objects i and j for variable t is invalid due to lack of data, otherwise $\delta_{ijt} = 1$. For binary variables $w_{ijt} = 1$, for nominal variables $w_{ijt} = x_{it} - x_{jt}$ if $x_{it} \neq x_{jt}$ and $w_{ijt} = 1$ if $x_{it} = x_{jt}$, for continuous variables $w_{ijt} = max\{x_{it}\} - min\{x_{jt}\}$ and for ordinal variables $w_{ijt} = max\{r_t\} - min\{r_t\}$.

Similarly to the Wishart distance, the Podani distance has similar properties to the Gower distance such that the same assumptions apply, missing values may be included and non-equal contribution of variables may be defined.

3.2 Generalization of the Gower's Coefficient with Weights

A modification that can be applied to the Gower's Coefficient involves adding a weighting method on the different attributes for when we consider a certain variable or group of variables to be more important than other.

One way to achieve this is by simply multiplying the numerator and denominator by the weight w given to the variable t the distance of which is being calculated. That gives a different weight w to each of the variables. The dissimilarity formula takes the following form:

$$d_{G,ij} = \frac{\sum_{t=1}^{p} \delta_{ijt} d_{ijt} w_t}{\sum_{t=1}^{p} \delta_{ijt} w_t} \tag{3.13}$$

The second way of using weights is by considering the difference between values of two different data records (Gower, 1971). This case is applied if we want to give more importance to

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

CC BY: Attribution alone 4.0
https://creativecommons.org/licenses/by/4.0/ https://doi.org/10.26219/heal.aueb.5832

a comparison whose values on both data records, x_{it} and x_{jt}, have an interesting property when being compared. In the formula we can state how much weight to give a certain comparison by changing w_t to a function of w that is dependent on x_{it} and x_{jt}. Specifically, the weight can be changed into $w(x_{it},x_{jt})$, which is a function that takes two values in the domain of the variable t and returns a value in the positive real domain, which is the corresponding weight. Then, the similarity coefficient will take the following form.

$$d_{G,ij} = \frac{\sum_{t=1}^{p} \delta_{ijt} d_{ijt} w(x_{it}, x_{jt})}{\sum_{t=1}^{p} \delta_{ijt} w(x_{it}, x_{jt})} \tag{3.14}$$

For both options, by adding the weights, the main properties of the dissimilarity formula remain, one being that the result of the measure calculation is a value between 0 and 1 (Sanchez,2021).

The first version with no weights corresponds to setting $w_t = 1$ for all variables. Gower (1971) mentions that setting rational weights is a difficult tasks, but it should be thought that assigning equal weights ($w_t = 1$) to all variables results in a balanced contribution of the variables to the distance in total (D'Orazio, 2021).

Although Gower's metric allows the inclusion of weights, not much literature concerning the selection process of the weights exists that defines the appropriate method. It is often mentioned that a selection of good weights may be achieved by using information known about the attributes, but a clear process is not available (Petchey & Gaston, 2009). It is most often that the study of Gower's coefficient is performed by using weights set to one (van den Hoven, 2015) and while methods have been proposed to add weights to the mixed type of data, they are computationally expensive.

Even so, the selection of weights can be a useful tool for gaining insight into the importance of different variables. Van den Hoven (2015) has mentioned that a proper weights selection procedure has the potential to improve clustering results. This occurs as attributes that are able to separate the data into clusters may be given higher weights, while attributes that produce poorer separation may be given a lower weight. This continues to be a difficult task though, as prior information about the attribute characteristics is necessary for the process. As an example of this, let us suppose this procedure results in some variables receiving a weight of 0, beneficially removing them from the process. Therefore, the variable would not need to be included in the clustering algorithm. Generally, the use of weights may assist in making a distinction between important and unimportant variables, and thus may improve the clustering quality by removing

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

CC BY: Attribution alone 4.0
https://creativecommons.org/licenses/by/4.0/
https://doi.org/10.26219/heal.aueb.5832

the noise from unessential variables. Various methods have been suggested for selecting weights, but there is no general way to choose the correct values.

Many have attempted to build a weighting system for Gower's measure since there is potential for enhancing clustering in selecting correct weights. For example, weights might be used to balance the effect of categorical and numeric variables (Chae, et al., 2006). Another approach is to base the weights on information from the variables (Montanari & Mignani, 1994). Both of these weighting techniques, however, are based on assumptions about procedures that may or may not work effectively for all datasets. This might lead to a good performance on some datasets but could also could lead to a very poor performance on other datasets.

Friedman and Meulman (2004) describe two algorithms for Clustering Objects on Subsets of Attributes (COSA), based on previously defined distance measures. The methods use a Gower-like distance metric, and suggests an interesting process for choosing attribute weights, but is explicitly based on the assumption that setting each weight to $1/p$ is equivalent to giving the same influence to all attributes regardless of data type, where p denotes the number of variables of any type. By selecting weights and cluster memberships simultaneously, Friedman and Muelman aim to optimise a criterion based on aggregated sums of Euclidean distances within clusters.

Van den Hoven (2015) developed a weighting scheme for Gower's metric that is not based on or at least relies less on external information or assumptions such as other methods. Their method suggested an optimal set of weights for Gower's metric through random search with a Quasi-Newton algorithm (Nocedal & Wright, 1999). These algorithms are used in order to find the minimum of a function by using the first and second order derivative. For this task, the Quasi-Newton method was used to optimise the weight vector in order to maximise the target function, which was the correlation between two distance matrices (Rocha, et al., 2011). The method that was used is called the limited memory Broyden - Fletcher - Goldfarb - Shanno algorithm with bounds (L-BFGS-B) (Nocedal & Wright, 1999).

3.3 Other Similarity Measures

Apart from Gower's similarity measure and its modification, other similarity methods have also been proposed in the literature for clustering mixed data. Some of these methods are reviewed within this chapter.

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

CC BY: Attribution alone 4.0
https://creativecommons.org/licenses/by/4.0/
https://doi.org/10.26219/heal.aueb.5832

3.3.1 Laflin's General Similarity Coefficient

Suppose N_k is the number of attributes for each different type and s_k is the similarity measure calculated for the different types of variables. For the calculation of the similarity measure between object i and j for a variable t, distance measures that have already been established for various attribute types are used and then transformed into similarity by the formula $s_k = \frac{1}{1+d_k}$, where d_k is the distance measure for the data of the same variable type. Specifically, the Euclidean distance is used for the numeric variables, while for categorical variables are calculated using the formula $d(i,j) = \frac{p-m}{p}$, where p is the total number of categorical variables and m is the number of categorical variables for which i and j have the same value.

For example, N_1, N_2, N_3, N_4 are the number of binary, numeric, ordinal and nominal attributes in a dataset and s_1, s_2, s_3, and s_4 are the similarity measures calculated for each pair of objects (Azam & Victor, 2011). The formula is dependent on the number of different variable types. The similarity values should be scaled so that they take the values between 0 and 1. Laflin's coefficient (Laflin, 1998) for the mixed set of attributes is defined as:

$$s_{ij} = \frac{\sum_{k=1}^{p} N_k s_{ijk}}{\sum_{k=1}^{p} N_k} \tag{3.15}$$

By using this formula, each attribute makes an equal contribution to the measure of similarity (Laflin, 1998).

Gower's coefficient and Laflin's coefficient both have the same denominator, but differ in terms of the numerator. Binary and nominal attributes are handled similarly by both, while a difference in the equations occurs from the functions selected for handling the numeric attributes. Due to this, it is indicated that the performances of the original Gower's similarity coefficient and the Laflin's similarity coefficient are very similar. Specifically, this is the case under certain conditions, where the dataset does not include asymmetric binary variables, and the distance measures for each of the variables are the same. In the experiments performed by Azam and Viktor (2011), it is observed through the evaluation measures used, that the original Gower coefficient, performs on average better than Laflin's coefficient. Better results are specifically obtained when using the Gower's coefficient in datasets containing more numeric variables than nominal or binary variables. It is suggested that Laflin's coefficient may be improved by incorporating a different function for the numeric variables.

CC BY: Attribution alone 4.0
https://creativecommons.org/licenses/by/4.0/
https://doi.org/10.26219/heal.aueb.5832

3.3.2 Huang Distance Measure (K-prototypes)

The k-prototypes algorithm, that was introduced by Huang (1997), is considered to be a popular clustering strategy that uses a mixed data distance technique. The distance measure between two mixed type variables, which are described by a total of p numeric attributes and $m - p$ categorical attributes, is described by the following formula:

$$d_{G,ij} = \sum_{t=1}^{p} \left(x_{it} - x_{jt}\right)^2 + \gamma \sum_{t=p+1}^{m} \delta\left(x_{it}, x_{jt}\right) \tag{3.16}$$

where the first term is the squared Euclidean distance measure on the numeric attributes and the second term is the simple matching dissimilarity measure on the categorical attributes. The weight γ is used to remove the preference of either type of attribute (Huang,1998).

The simple matching similarity measure for categorical variables is denoted as follows.

$$\delta\left(x_{it}, x_{jt}\right) = \begin{cases} 0, & x_{it} = x_{jt} \\ 1, & x_{it} \neq x_{jt} \end{cases} \tag{3.17}$$

In contrary to Gower's similarity measure, the K-prototypes distance specified in the paragraph, does not allow variable-specific weights. Instead, a single weight is assigned to weight the categorical contribution to the distance. This weight, as happens in the Gower's similarity also, is difficult or impossible to choose optimally. In addition, the K-prototypes distance suffers from the problem of not considering the similarity between categorical values optimally, as some categories may be more similar than others, something that is not taken into account with the simple matching measure. Thereby, clustering with this method could not correctly uncover the similarity structure of the data involving categorical values with different levels of similarity.

3.3.3 Ahmad and Dey Distance Measure (K-means)

Ahmad and Dey (2007) have proposed a fuzzy prototype k-means algorithm for a mixed type dataset, that is an extension of Huang's distance. This distance measure uses the squared Euclidean distance for continuous variables, while the distance for categorical variables is calculated as a function of the distribution and co-occurrence of the data objects with other variables, and is given by the following formula:

CC BY: Attribution alone 4.0
https://creativecommons.org/licenses/by/4.0/
https://doi.org/10.26219/heal.aueb.5832

$$d_{ij} = \sum_{k=1}^{P_n} \left(x_{ik} - x_{jk}\right)^2 + \sum_{k=1}^{P_c} \delta_c\left(x_{ik}; x_{jk}\right) \tag{3.18}$$

where $\delta_c\left(x_{ik}; x_{jk}\right)$ is the co-occurrence distance.

The co-occurrence distance is the measure used for calculating the distance between the categorical variables and it is based on a probabilistic approach. Suppose x and y are two categorical values of an object A_i. In order to find the distance between x and y, the method considers the overall distribution of x and y in the data set along with their co-occurrence with values of other variables. Then the probability of occurrence of x and y values with other categorical attribute values A_j of the dataset is computed. Let w denote a subset of values of A_j and $\overline{w}$ the complementary set of values occurring for variable A_j. Let $\mathrm{P}(w|x)$ denote the conditional probability that an object having value x for A_i, has a value belonging to w for A_j, and similarly $\mathrm{P}(\overline{w}|y)$. Therefore, the distance between x and y of A_i, with respect to A_j and a specific subset w, is found with the following formula:

$$\delta^{ij}(x, y) = P_i(w|x) + P_i(\overline{w}|y) - 1 \tag{3.19}$$

where x is the subset w of values of A_i that maximizes the quantity $P_i(w|x) + P_i(\overline{w}|y)$. The distance between x and y is computed with respect to every other variable and the average value of distances will be the overall distance between x and y in the data set.

By using the co-occurrence distance for the categorical variables, Ahmad and Dey suggest that the distribution of the values is considered in the distance calculation. They mention that the distance between categorical variables cannot be binary, as values which co-occur in the same group should be thought of as more similar than values that co-occur in different groups. This is not incorporated in most categorical value distances (Ahmad & Dey, 2007; Bishnoi & Hooda, 2020).

As previously mentioned, Huang's distance measure used a binary valued distance for computing the distance between categorical variables. In addition, the contribution of the categorical variables to the overall distance was weighted by a user defined parameter. In Ahmad and Dey's distance measure the contribution of a categorical variables is ingrained in the distance measure and is a function of the value co-occurrence. Therefore, the user is not obliged to specify the weight for the computation as the weighing values were extracted from

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

CC BY: Attribution alone 4.0 https://doi.org/10.26219/heal.aueb.5832
https://creativecommons.org/licenses/by/4.0/

the variable value distributions within the data. In addition, the binary and categorical variables are not separable such that the co-occurrence distance is based on the combination of these two classes (Bishnoi & Hooda, 2020).

3.3.4 Harikumar & PV Distance Measure

Harikumar and PV (2015) have suggested an alternative generalized distance function in the form of a triplet. This includes three different distance measures for numeric, categorical and binary data types. Therefore, the proposed similarity measure has three distinct components, one for dealing with numeric attributes, second for dealing with categorical attributes and third for handling binary attributes, as in the following formula (Harikumar & PV, 2015):

$$d_{ij} = \sum_{r=1}^{P_n} \left|x_{ir} - x_{jr}\right| + \sum_{s=1}^{P_c} \delta_c\left(x_{is}, x_{js}\right) + \sum_{t=1}^{p_b} \delta_b\left(x_{it}, x_{jt}\right) \tag{3.20}$$

where P_n is the number of numeric attributes, P_c the number of categorical attributes and P_b the number of binary variables (i.e., Total attributes $P = P_n + P_c + P_b$), $\delta_c\left(x_{is}, x_{js}\right)$ is the co-occurrence distance and $\delta_b\left(x_{it}, x_{jt}\right)$ is the Hamming distance.

The first component is the distance between two numeric objects, that is calculated with the help of norms (Fayyad & Smyth,1996). For an object x_i and x_j, with P_n numeric attributes, $L_k$is defined as

$$L_k = \left(\sum_{r=1}^{P_n} \left|x_{ir} - x_{jr}\right|\right)^{\frac{1}{k}} \tag{3.21}$$

When $k = 1$, it is L_1 norm, when $k = 2$, it is L_2 norm etc.. The L_1 norm is the absolute difference between each attribute of two objects. According to Harimukar and PV, L_1 norm is flexible, robust and resistant to outliers and therefore it is used to calculate the distance.

The co-occurrence distance is the measure used for calculating the distance between the categorical variables and it is based on a probabilistic approach, as described in the previous section.

Before applying the distance measure for similarity on the data, the numeric attributes are normalized as in the following:

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

CC BY: Attribution alone 4.0
https://creativecommons.org/licenses/by/4.0/
https://doi.org/10.26219/heal.aueb.5832

$$x_{new} = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{3.22}$$

Lastly, the simple matching is used for the binary variables. According to this distance, two Boolean attribute values x and y that is $\delta(x, y) = 0$ for $x = y$ and $\delta(x, y) = 1$ for $x \neq y$ (Harikumar & PV, 2015).

For this distance measure, categorical variables were treated separately from binary variables, due to the variations in their probability distributions, unlike Ahmad and Dey, who used the same measure for binary and categorical variables. The experiments performed by Harikumar and PV (2015) showed the good performance of their distance measure, while it also was shown to surpass the performance of the Ahmad and Dey distance measure.

3.3.5 Modha and Spangler (convex K-means)

Modha and Spangler (2003) created a method that is very similar to the k-prototypes method, which reduces to a single weight that determines the relative contribution of continuous and categorical variables by using a weighted combination of squared Euclidean distance and cosine distance for the continuous and categorical variables, respectively (Foss & Markatou, 2018).

Instead of a user-specified weight, their method looks for an optimal weight to minimize the ratio of the within and between cluster weighted distances when the number of clusters is given. This way their method calculates an appropriate weight w_k that scales the relative contribution of the interval and categorical scale variables.

The numeric attributes are standardized based on the mean and standard deviance, and the distance is calculated by using the squared-Euclidean distance. The categorical variables are represented by a binary vector using 1-of-v encoding, where v is the number of attributes, and the distance is calculated by using the cosine distance. The optimal weight distance combines the weighted distances of the two data types as follows (Yang, 2012).

$$d(x, y) = (1 - w) \sum_{i \in A_r} \frac{(a_i(x) - a_i(y))^2}{\operatorname{var}(V_{a_i})} + w \sum_{i \in A_c} D^i(x, y) \tag{3.23}$$

The weight in the Modha-Spangler algorithm is identified through a brute-force search to cluster repeatedly over a range of user-specified values. Modha-Spangler clustering is considered a useful method as it has the ability to balance the contribution of continuous and

CC BY: Attribution alone 4.0
https://creativecommons.org/licenses/by/4.0/
https://doi.org/10.26219/heal.aueb.5832

categorical variables at a high standard. An issue to this technique arises as the single weight does not allow individual variables to be weighted differently. For instance, when categorical variables of a dataset vary in strength of association with the cluster structure, all variables need to be up or down weighted uniformly (Foss, Markatou, Ray, & Heching, 2016).

3.3.6 Similarity Based Agglomerative Clustering - Li & Biswas

Li and Biswas (1998) presented the Similarity Based Agglomerative Clustering (SBAC) algorithm which is based on Goodall similarity metric. The Goodall similarity measure (1966) is based on the concept that two species are closer if they have rarer characteristics in common, and defines a unified framework for dealing with mixed data. Specifically, a pair of objects (i, j) is thought to be more similar than another pair of objects (l, m) if the objects i and j show a greater match in the attribute values that are more uncommon in the population (Li & Biswas, 1998). This is included in the similarity measure by weighing features by the frequency of occurrence of their attribute-values in the dataset and therefore leading uncommon attribute values to make more significant contributions to the overall similarity between the objects.

Suppose A is set of attributes, where A_r is a set of numeric attributes and A_c is a set of categorical attributes, n the number of objects. There exists a function between the set of objects and each attribute such that $a_p: U \rightarrow V_{a_p}$ for any $a_p \in A(p = 1,2, \ldots, m)$, i.e., $a_p(x_i) \in V_{a_p}(i = 1,2, \ldots, n), x_i \in U$, where V_{a_p} is called the domain of a_p. $a_p(x_i)$ is the value of object x_i on attribute a_p. The distance between two objects x and y on a_p is defined as $D^p(x, y)$. The total distance between two objects x and y is defined as $d(x, y)$.

The distance for non-identical nominal values is one as $D^i(x, y) = 1, a_i(x) \neq a_i(y), a_i \in A_c$. For a pair of identical nominal values, the distance is the sum of the possibilities of choosing an identical value pair whose value is equally or more similar to the pair in question, that is having lower or equal frequency. This is as follows when $s = a_i(x) = a_i(y), a_i \in A_c$.

$$D^i(x, y) = \sum_{r \in V_{a_i},\ \text{freq}\,(r) \leq \text{freq}\,(s)} \frac{\text{freq}\,(r)(\,\text{freq}\,(r) - 1)}{n(n-1)} \tag{3.24}$$

The distance between identical numeric values is calculated as their nominal component using the above equation. To calculate distance of two different numeric values, the domain is divided into successive segments by the unique values of a numeric feature. Frequency is

CC BY: Attribution alone 4.0
https://creativecommons.org/licenses/by/4.0/
https://doi.org/10.26219/heal.aueb.5832

counted in every interval first. Then, sum the possibilities in a smaller range or equal-width range (l, m) but with less or equal frequency. Given $s = a_i(x), t = a_i(y), a_i(x) \neq a_i(y), a_i \in A_r$, the distance on a numeric attribute A_i is:

$$D^i(x, y) = \sum_{r \in V_{a_i}} \frac{\text{freq}(r)\,(\text{ freq }(r) - 1)}{n(n-1)} + \sum_{\substack{|m-l|<|t-s| \vee \{|m-l|=|t-s|, \\ freq(|m-l|) \leq freq(|t-s|)\}}} \frac{2\text{freq }(l)(\text{ freq }(m) - 1)}{n(n-1)} \tag{3.25}$$

When having calculated the distances for each attribute, the χ^2 transformation is used to get corresponding chi-square values. The sum of these values is distributed as χ^2 with the degree of freedom two times of the number of attributes. The probability of this sum is the Goodall distance of a pair of objects (Yang, 2012).

3.3.7 TreeClust Dissimilarity

A dissimilarity measure named TreeClust was introduced by Buttrey (2006), that utilizes classification and regression trees to produce the measure. The main concept of this measure is the following: Two objects are similar if they are usually found in the same leaves of the classification or regression tree. In contrast to the clustering method, a regression or classification tree uses a response variable and therefore p trees need to be created, one for each variable. Each variable will be used as the response variable whilst the other variables will be used as predictive variables. The attributes that are categorical are shown by classification trees whereas the numeric attributes will be shown as regression trees (Buttrey, 2016).

Once the trees are created, the dissimilarity between two objects is calculated. This measure is found as the number of trees in which the two objects fall into different leaves. Let $L_t(i)$ be the label of the leaf of the t^{th} tree into which the i^{th} object falls. Then the TreeClust dissimilarity between objects i and j is described as:

$$d_{ij} = \sum_{t=1}^{p} \delta\left(x_{it}, x_{jt}\right) \tag{3.26}$$

where

$$\delta\left(x_{it}, x_{jt}\right) = \begin{cases} 0, & L(x_{it}) = L(x_{jt}) \\ 1, & L(x_{it}) \neq L(x_{jt}) \end{cases} \tag{3.27}$$

CC BY: Attribution alone 4.0
https://creativecommons.org/licenses/by/4.0/
https://doi.org/10.26219/heal.aueb.5832

Therefore, for tree t, two observations have a distance of 1 if they fall on different leaves, while if observations fall on the same leaf, they have a distance of 0 (Buttrey & Whitaker, 2016).

Buttrey and Whitaker (2016) mention that this distance measure holds the advantages of classification and regression trees. Categorical and numeric variables are combined and weighted in an obvious way, and as they are local-type models, they are resistant to outliers. In addition, the trees are easily built, even in the presence of missing values.

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

CC BY: Attribution alone 4.0
https://creativecommons.org/licenses/by/4.0/ https://doi.org/10.26219/heal.aueb.5832

CHAPTER 4

Simulation Study

4.1 Simulation Setup

In order to evaluate the similarity measures that were presented in Chapter 3, and to determine how each measure performs in clustering data having a variety of structures, experiments on simulated datasets were included within this thesis. Simulated data sets are frequently used to numerically evaluate the performance of clustering methods, as they have known cluster structures so that we can evaluate the performance of a clustering method by checking the agreement between the true partition and the partition found by the clustering method. The noise (e.g. noisy variables, outliers, missing values, and measurement errors, etc.) can also be controlled, while multiple replicates can also be produced. Furthermore, they can identify the situations in which a clustering method performs well or not.

The purpose of these experiments was to see whether the clustering results were different and how such a difference would depend on the underlying cluster structure, such as a difficult structure defined by the number of categories of the clustered variables or the ratio of quantitative and categorical variables in the dataset. Therefore, a first experiment was conducted in which the dataset characteristics were modified and tested, while the second experiment used a dataset that included a high number of total variables. With the results produced from the experiments it is feasible to decide in which case a certain dissimilarity metric is more beneficial compared to others, and which is identified by an overall higher performance.

The data generation method is similar to that mentioned in Sulc et al. (2017). For the dataset simulation of the first experiment performed in this thesis, a total of 270 datasets were generated, whilst for the second experiment 20 datasets were tested. The total number of objects in the produced datasets was not fixed, but was randomly selected to be between 300 and 700 objects, while the total number of clusters in all was set to four. For the first experiment, each specific combination of characteristics was reproduced twice. In our experiments, all the algorithms were implemented in R language.

For the implementation in R, the datasets were generated by using the *genRandomClust()* function from the *clusterGeneration* R package (Qiu & Joe, 2015). This can be used to

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

CC BY: Attribution alone 4.0
https://creativecommons.org/licenses/by/4.0/
https://doi.org/10.26219/heal.aueb.5832

generate cluster datasets in which the separation between any cluster and its nearest neighbouring cluster can be set to a specified value, while the covariance matrices of clusters can also have arbitrary diameters, shapes and orientations. The Qiu and Joe algorithm (Qiu & Joe, 2006) is limited to generating only numerical attribute datasets. Therefore, to be able to produce categorical variables within the datasets, a numerical variable is categorized through a quantile categorization. This categorization of numerical variables to categorical variables, is achieved by dividing the numerical variable into c- number of classes with $100/c$ % quantile as the cut point. Thus, a binary variable is achieved when a median of the numerical variable is applied as a cut-off point. The generated datasets had a predefined number of clusters which was set to four, and the correct grouping of each object in all datasets was known.

The mixed variables distances were directly implemented in the R environment through a variety of packages. Specifically, for the calculation of Gower's General Dissimilarity Coefficient, the Inter-Quartile and Inter-Decile Gower's Measure Modification, the Modified Kernel Density Estimation Distance, the Modified k-nn Distance and for the Modification Based on 'Conditional Distance', the *'StatMatch'* package was used (D'Orazio, 2022). In addition, Laflin's Distance is calculated by using the *'cluster'* R package and the Similarity Based Agglomerative Clustering Distance by the *'nomclust'* R package (Šulc et al., 2022; Maechler et al., 2013). The Inverse Occurrence Frequency Measure, the Lin Measure, and the Variable Entropy Measure were calculated by using a combination of the two previously mentioned packages, *'cluster'* and *'nomclust'*. The Wishart Distance, Podani Distance, Huang Distance, Harikumar-PV Distance and Ahmad and Dey Distance are found by using the *'kmed'* R package (Budiaji, 2019). Lastly, the Modha & Sprangler Distance was computed with the *'kamila'* R package, while the TreeClust Distance by using the *'treeclust'* R package (Foss & Markatou, 2018; Buttrey & Whitaker, 2015).

The symbolisms for the measures presented previously, that are used in this chapter for simplicity, are shown in *Table 4.1*.

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

CC BY: Attribution alone 4.0
https://creativecommons.org/licenses/by/4.0/
https://doi.org/10.26219/heal.aueb.5832

Measure Name	Symbolism
Gower's General Dissimilarity Coefficient	Gower
Inter-Quartile Gower's measure modification - D'Orazio	G_IQ
Inter-Decile Gower's measure modification - D'Orazio	G_ID
Modified Kernel Density Estimation Distance - D'Orazio	G_KDE
Modified k-nn Distance - D'Orazio	G_KNN
Modification Based on 'Conditional Distance' - D'Orazio	G_CD
Inverse Occurrence Frequency – Šulc, Z., Matejka, M., & Procházka, J.	G_IOF
Lin Measure - Šulc, Z., Matejka, M., & Procházka, J.	G_Lin
Variable Entropy Measure - Šulc, Z., Matejka, M., & Procházka, J.	G_VE
Wishart's Dissimilarity Measure	G_Wishart
Podani's Dissimilarity Measure	G_Podani
Laflin's General Similarity Coefficient	Laflin
Huang distance measure (K-prototypes)	Huang
Ahmad & Dey (K-means)	Harikumar-PV
Harikumar & PV distance measure	Ahmad & Dey
Modha & Spangler (convex K-means)	Modha & Sprangler
Similarity Based Agglomerative Clustering - Li & Biswas	SBAC
TreeClust Dissimilarity	TreeClust

Table 4.1 Measure symbolisms

As mentioned Hierarchical clustering is a clustering method that creates a hierarchy of clusters by using a similarity matrix (Jain & Dubes, 1988). The similarity matrix is constructed by determining the similarity of each pair of data points. In the experiments, Hierarchical Clustering was conducted on each of the produced datasets for the mentioned similarity measures in Chapter 3. The complete linkage method was used to obtain the results. The clustering process is performed by using the *hclust* command from the '*factoextra'* R package.

The resulting clusters were evaluated by the average values of the Rand indices, which in the first experiment were broken down by the minimal between cluster distance, difficulty structure and combinations of quantitative and categorical variables, while also compared as the average of all datasets as in the second experiment.

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

CC BY: Attribution alone 4.0
https://creativecommons.org/licenses/by/4.0/
https://doi.org/10.26219/heal.aueb.5832

4.2 First Experiment – Dataset Variation

4.2.1 Dataset Characteristics

For the first experiment, the simulated datasets are comprised by continuous and categorical variables and have three characteristics, modified between them, so that each dataset is different.

Firstly, they differ by their minimal between-cluster distance. The minimal between-cluster distance is the separation between a cluster and its nearest neighbouring cluster. This separation between a pair of clusters is measured by the separation index proposed in Qiu and Joe (2006), and is accomplished in the R package by setting *sepVal* to the equivalent distance. Theoretically, *sepVal* can take values between -1 and 1, and the closer it is to 1 the more separated the clusters are. Thus, the higher the distance the more clear the cut is between the clusters. For this experiment, the distance is categorised into low, middle and high depending on whether the minimal between-cluster distance is 0.1, 0.3 or 0.5 respectively. With the high distance of 0.5, most clusters do not cross each other in the datasets.

As a second characteristic, three levels of difficulty where stated. The levels were easy, medium and hard based on the numbers of categories of the categorical variables, as well as any feasible combinations of quantitative and categorical variable counts, excluding the case in which all variables are categorical and none are numeric. In the easy level, the categorical variables have two to four categories, in the medium level they have three to seven categories and in the difficult level they have eight to ten categories.

The third characteristic that was tested was the total number of variables that were included in the dataset. The total number of variables was changed between four, six and eight variables.

The number of clusters were predefined in all datasets to be a total of four, and each object's cluster was known. The total number of objects in the produced datasets was not fixed, but they varied from 300 to 700, and was chosen randomly. Each specific combination of characteristics was reproduced twice, to guarantee the robustness of the acquired results. Therefore, a total of 270 datasets were simulated (three minimal between-cluster distance levels x three levels of difficulty x fifteen combinations of numeric and categorical variables x reproduced twice).

CC BY: Attribution alone 4.0
https://creativecommons.org/licenses/by/4.0/
https://doi.org/10.26219/heal.aueb.5832

For the experiment, each of the characteristics was tested separately, by using all datasets simulated for that characteristic and then by averaging the results. For each characteristic tested in the study, the total number of datasets is the following.

- When testing the minimal between-cluster distance 45 datasets are tested for each distance level (three levels of difficulty x fifteen combinations of numeric and categorical variables). The number of iterations for the simulation study is 2.

- When testing the levels of difficulty also 45 datasets are tested for each difficulty level (three minimal between-cluster distance x fifteen combinations of numeric and categorical variables). The number of iterations for the simulation study is 2.

- When testing the number and combinations of variables, 27 datasets are tested for four variables (three minimal between-cluster distance levels x three levels of difficulty x three combinations of numeric and categorical variables), 45 datasets for six variables (three minimal between-cluster distance levels x three levels of difficulty x five combinations of numeric and categorical variables) and 63 for eight variables (three minimal between-cluster distance levels x three levels of difficulty x seven combinations of numeric and categorical variables). The number of iterations for the simulation study is 2 for each setting.

A summary of the dataset characteristics and their combinations, is shown in *Table 4.2*. The first column shows the number of variables, the second column shows the number of all categorical variables tested in each case. Each number of variable is combined with each number of categorical variables, difficulty level and minimal between-cluster distance and therefore producing a total of 270 datasets. The total number of objects that were produced in all datasets for the 1st and 2nd repetition of each combination of characteristics is also shown in the last two columns of *Table 4.2*. This varied between the first and second repetition and number of variables, but remained stable for the difficulty level and minimal between-cluster distance, as shown in the Table.

CC BY: Attribution alone 4.0
https://creativecommons.org/licenses/by/4.0/ https://doi.org/10.26219/heal.aueb.5832

# of variables	*#of cat. var*	*Difficulty*	*Distance*	*# of Objects 1st repetition*	*# of Objects 2nd repetition*
4	1, 2, 3	Easy	0.1	421, 638, 460	465, 342, 509
6	1, 2, 3, 4, 5	Easy	0.1	421, 486, 472, 431, 558	494, 443, 505, 499, 572
8	1, 2, 3, 4, 5, 6, 7	Easy	0.1	421, 465, 587, 413, 477, 579, 588	503, 429, 551, 635, 532, 549, 566
4	1, 2, 3	Medium	0.1	421, 638, 460	465, 342, 509
6	1, 2, 3, 4, 5	Medium	0.1	421, 486, 472, 431, 558	494, 443, 505, 499, 572
8	1, 2, 3, 4, 5, 6, 7	Medium	0.1	421, 465, 587, 413, 477, 579, 588	503, 429, 551, 635, 532, 549, 566
4	1, 2, 3	Hard	0.1	421, 638, 460	465, 342, 509
6	1, 2, 3, 4, 5	Hard	0.1	421, 486, 472, 431, 558	494, 443, 505, 499, 572
8	1, 2, 3, 4, 5, 6, 7	Hard	0.1	421, 465, 587, 413, 477, 579, 588	503, 429, 551, 635, 532, 549, 566
4	1, 2, 3	Easy	0.3	421, 638, 460	465, 342, 509
6	1, 2, 3, 4, 5	Easy	0.3	421, 486, 472, 431, 558	494, 443, 505, 499, 572
8	1, 2, 3, 4, 5, 6, 7	Easy	0.3	421, 465, 587, 413, 477, 579, 588	503, 429, 551, 635, 532, 549, 566
4	1, 2, 3	Medium	0.3	421, 638, 460	465, 342, 509
6	1, 2, 3, 4, 5	Medium	0.3	421, 486, 472, 431, 558	494, 443, 505, 499, 572
8	1, 2, 3, 4, 5, 6, 7	Medium	0.3	421, 465, 587, 413, 477, 579, 588	503, 429, 551, 635, 532, 549, 566
4	1, 2, 3	Hard	0.3	421, 638, 460	465, 342, 509
6	1, 2, 3, 4, 5	Hard	0.3	421, 486, 472, 431, 558	494, 443, 505, 499, 572
8	1, 2, 3, 4, 5, 6, 7	Hard	0.3	421, 465, 587, 413, 477, 579, 588	503, 429, 551, 635, 532, 549, 566
4	1, 2, 3	Easy	0.5	421, 638, 460	465, 342, 509
6	1, 2, 3, 4, 5	Easy	0.5	421, 486, 472, 431, 558	494, 443, 505, 499, 572
8	1, 2, 3, 4, 5, 6, 7	Easy	0.5	421, 465, 587, 413, 477, 579, 588	503, 429, 551, 635, 532, 549, 566
4	1, 2, 3	Medium	0.5	421, 638, 460	465, 342, 509
6	1, 2, 3, 4, 5	Medium	0.5	421, 486, 472, 431, 558	494, 443, 505, 499, 572
8	1, 2, 3, 4, 5, 6, 7	Medium	0.5	421, 465, 587, 413, 477, 579, 588	503, 429, 551, 635, 532, 549, 566
4	1, 2, 3	Hard	0.5	421, 638, 460	465, 342, 509
6	1, 2, 3, 4, 5	Hard	0.5	421, 486, 472, 431, 558	494, 443, 505, 499, 572
8	1, 2, 3, 4, 5, 6, 7	Hard	0.5	421, 465, 587, 413, 477, 579, 588	503, 429, 551, 635, 532, 549, 566

Table 4.2 Dataset characteristics and combinations

For example, the first dataset produced, has a total of 4 variables, one of which is categorical and three are numeric. The dataset difficulty is set to be easy, meaning that the number of categories of the categorical variable are randomly chosen to be either two, three or four. The minimal between-cluster distance is set to 0.1. The total number of objects for this example of characteristic combinations is 421 for the first repetition and 465 for the second repetition. Some descriptive statistics for the first repetition of the dataset with this combination of characteristics are shown in *Table 4.3*.

CC BY: Attribution alone 4.0
https://creativecommons.org/licenses/by/4.0/
https://doi.org/10.26219/heal.aueb.5832

Variable Type	***1st variable***	***2nd variable***	***3rd variable***	***4th variable***
Categorical	Yes			
# of Categories	3	-	-	-
# of objects in 1st category	141	-	-	-
# of objects in 2nd category	140	-	-	-
# of objects in 3rd category	140	-	-	-
Numeric		Yes	Yes	Yes
Minimum	-	-18.71	-16.70	-11.56
1st Quartile	-	-3.84	-10.25	0.63
Median	-	0.19	-5.31	3.02
Mean	-	-0.67	-5.00	3.06
3rd Quartile	-	2.72	0.18	5.76
Maximum	-	9.32	7.08	17.83

Table 4.3 Descriptive statistics for example dataset.

The box plots for the categorical variable against the numeric variables for the example dataset is shown in *Figure 4.1*.

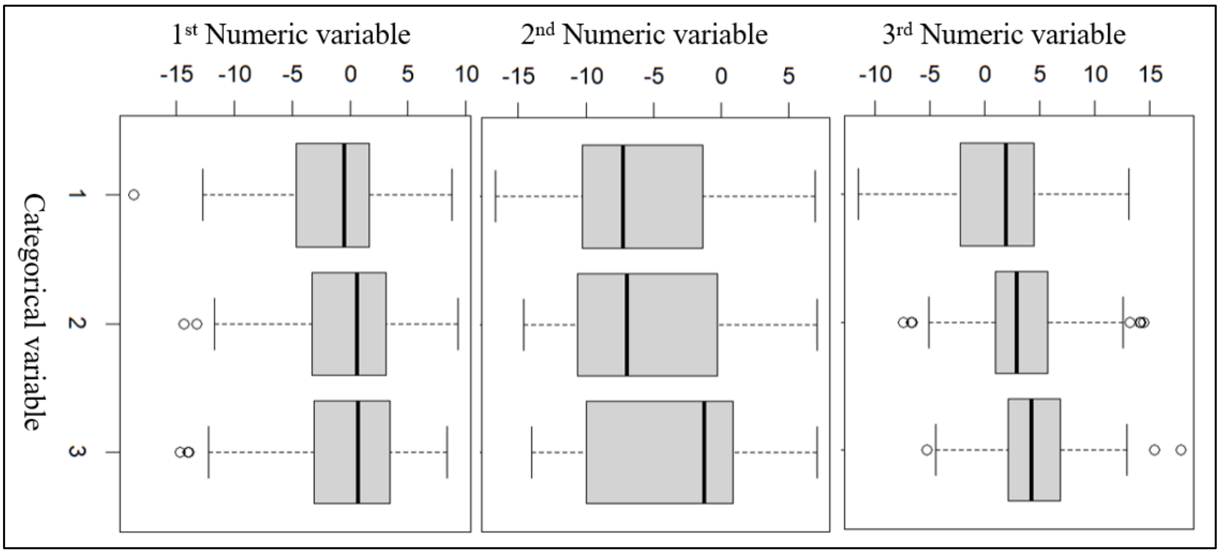

Figure 4.1 Box plot for example dataset

In addition , a PCA graph for the example dataset is shown in *Figure 4.2*. PCA reduces the number of dimensions by constructing principal components. In this Figure, the correct cluster for each object is shown by its colouring while a total of four clusters are seen for the data.

CC BY: Attribution alone 4.0
https://creativecommons.org/licenses/by/4.0/
https://doi.org/10.26219/heal.aueb.5832

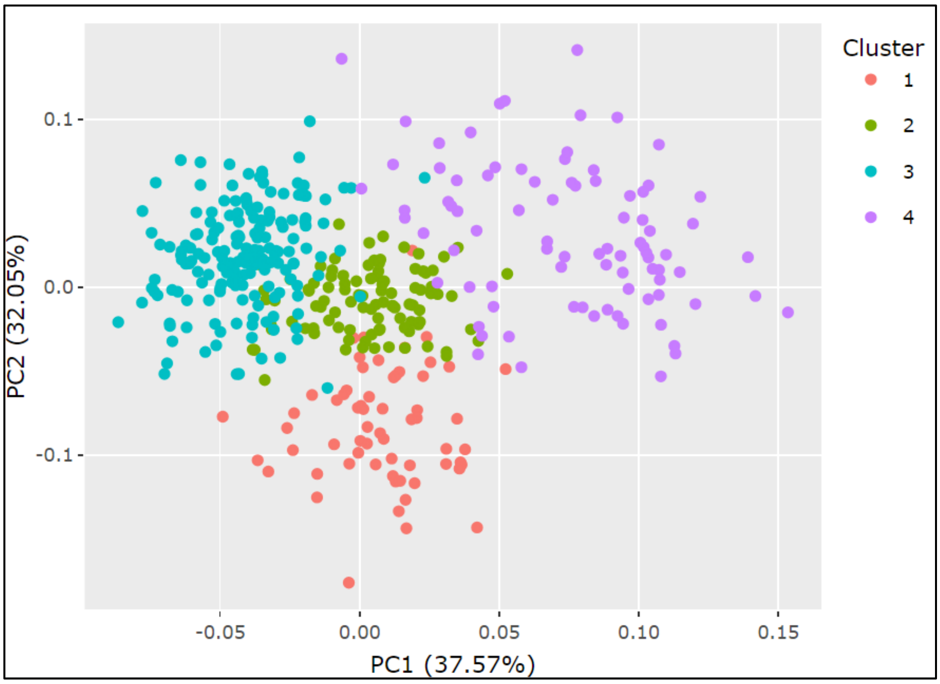

Figure 4.2 PCA plot of example dataset

4.2.2 Results for Dataset Difficulty Variation

Initially, the impact of the dataset difficulty was examined. The difficulty level is associated with the number of categories held by the categorical variables of the datasets. As mentioned, three difficulty levels are included, easy, medium and hard. The categorical variables contain two to four categories in the easy level, three to seven for the medium level and eight to ten in the hard level.

Each level is tested for 90 datasets, that include variation in the minimal between-cluster distance, the number of variables and the combination of numeric and categorical variables, as specified above.

The measures were then calculated on each of the datasets, and the average Rand index for each was found. *Table 4.4* depicts the average values of the Rand indices for the similarity measures shown in Chapter 3, with each being an average of 90 Rand indices, as they were calculated from a total of 90 datasets each.

CC BY: Attribution alone 4.0
https://creativecommons.org/licenses/by/4.0/
https://doi.org/10.26219/heal.aueb.5832

Difficulty	*Gower*	*G_IQ*	*G_ID*	*G_KDE*	*G_KNN*	*G_CD*
Easy	0.775	0.732	0.783	0.483	0.674	0.467
Medium	0.721	0.621	0.696	0.323	0.550	0.436
Hard	0.710	0.588	0.678	0.310	0.509	0.425
	G_IOF	*G_Lin*	*G_VE*	*G_Wishart*	*G_Podani*	*Laflin*
Easy	0.758	0.751	0.751	0.735	0.747	0.741
Medium	0.692	0.637	0.587	0.723	0.690	0.661
Hard	0.670	0.648	0.602	0.725	0.674	0.655
	Huang	*Harikumar-PV*	*Ahmad & Dey*	*Modha & Sprangler*	*SBAC*	*TreeClust*
Easy	0.741	0.818	0.804	0.798	0.623	0.646
Medium	0.661	0.780	0.795	0.782	0.610	0.551
Hard	0.655	0.761	0.755	0.644	0.625	0.489

Table 4.4 Average value of Rand index by dataset difficulty

For all the included similarity measures, it is seen that as the difficulty of the dataset becomes stronger, the Rand index decreases. G_VE is an exception as its performance worsens from easy to medium, but has a slight increase from medium to hard.

For the easy datasets, Harikumar-PV and Ahmad & Dey similarity measures have a leading performance. For the medium difficulty datasets Harikumar-PV, Ahmad & Dey and Modha & Sprangler measures have the highest Rand indices. For the hard datasets the Rand index prefers Harikumar-PV, Ahmad & Dey, while Gowers coefficient, G_Wishart also has a high Rand index which is above 0.7.

The Gower's similarity, G_Wishart and SBAC are seen to not highly depend on the difficulty level, as the Rand index doesn't change significantly when moving between easy, medium and hard.

The variances of the above measures are shown in *Table 4.5*.

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

CC BY: Attribution alone 4.0
https://creativecommons.org/licenses/by/4.0/
https://doi.org/10.26219/heal.aueb.5832

Difficulty	*Gower*	*G_IQ*	*G_ID*	*G_KDE*	*G_KNN*	*G_CD*
Easy	0.008	0.024	0.020	0.031	0.010	0.051
Medium	0.013	0.045	0.045	0.032	0.012	0.065
Hard	0.012	0.058	0.053	0.042	0.015	0.068
	G_IOF	*G_Lin*	*G_VE*	*G_Wishart*	*G_Podani*	*Laflin*
Easy	0.007	0.010	0.012	0.019	0.007	0.007
Medium	0.006	0.023	0.021	0.020	0.010	0.010
Hard	0.008	0.029	0.020	0.028	0.017	0.012
	Huang	*Harikumar-PV*	*Ahmad & Dey*	*Modha & Sprangler*	*SBAC*	*TreeClust*
Easy	0.029	0.032	0.033	0.014	0.017	0.009
Medium	0.030	0.033	0.029	0.016	0.015	0.010
Hard	0.034	0.037	0.034	0.015	0.020	0.012

Table 4.5 Variance of Rand index by dataset difficulty

4.2.3 Results for Minimal Between-Cluster Distances Variations

The measures were also examined by a minimal between-cluster distance of 0.1, 0.3 and 0.5 which are described as low, middle and high distance, respectively. Similarly as in the dataset difficulty testing, each level of minimal between-cluster distance is tested for 90 datasets, that include variation in the level of difficulty, the number of variables and the combination of numeric and categorical variables. The measures are then calculated on each of the datasets, and the average Rand index for each is found. The Rand index for each of these is shown in *Table 4.6*, with each being an average of 90 Rand indices, as they were calculated from a total of 90 datasets each.

As the minimal between-cluster distance increases from low to high, the Rand indices of all similarity measures increases, apart from Modha & Sprangler and SBAC. For low distance datasets, the Huang, the Harikumar-PV and the TreeClust measures have the highest performance. For middle distance datasets also the previous measures are preferred. For high distance datasets, Gowers coefficient, G_ID and Ahmad & Dey are also included with the previous measures as top performance measures.

CC BY: Attribution alone 4.0
https://creativecommons.org/licenses/by/4.0/
https://doi.org/10.26219/heal.aueb.5832

Distance	*Gower*	*G_IQ*	*G_ID*	*G_KDE*	*G_KNN*	*G_CD*
Low	0.655	0.526	0.576	0.348	0.547	0.429
Middle	0.723	0.633	0.743	0.369	0.578	0.444
High	0.828	0.783	0.839	0.399	0.609	0.450
	G_IOF	*G_Lin*	*G_VE*	*G_Wishart*	*G_Podani*	*Laflin*
Low	0.657	0.582	0.573	0.683	0.643	0.644
Middle	0.703	0.677	0.660	0.721	0.698	0.690
High	0.759	0.776	0.752	0.783	0.788	0.732
	Huang	*Harikumar-PV*	*Ahmad & Dey*	*Modha & Sprangler*	*SBAC*	*TreeClust*
Low	0.717	0.718	0.698	0.629	0.682	0.722
Middle	0.793	0.788	0.770	0.544	0.591	0.781
High	0.850	0.848	0.828	0.450	0.513	0.854

Table 4.6 Average values of Rand index by minimal cluster distances.

The variances of the above measures are shown in *Table 4.7*.

Distance	*Gower*	*G_IQ*	*G_ID*	*G_KDE*	*G_KNN*	*G_CD*
Low	0.004	0.027	0.026	0.009	0.011	0.056
Middle	0.007	0.043	0.031	0.014	0.012	0.066
High	0.009	0.044	0.029	0.025	0.017	0.074
	G_IOF	*G_Lin*	*G_VE*	*G_Wishart*	*G_Podani*	*Laflin*
Low	0.004	0.013	0.012	0.011	0.004	0.004
Middle	0.008	0.017	0.015	0.017	0.007	0.007
High	0.009	0.022	0.023	0.021	0.011	0.009
	Huang	*Harikumar-PV*	*Ahmad & Dey*	*Modha & Sprangler*	*SBAC*	*TreeClust*
Low	0.025	0.028	0.024	0.056	0.022	0.003
Middle	0.032	0.030	0.030	0.066	0.026	0.010
High	0.028	0.033	0.028	0.074	0.029	0.009

Table 4.7 Variance of Rand index by minimal cluster distances.

4.2.4 Results for the Combinations of Numeric and Categorical Variables

Lastly, the combination of different number of variables (4,6 and 8), that include numeric and categorical variables. All different number of categorical variables (except for the case of all numeric and all categorical variables) were examined. Specifically, for datasets with a total of four variables, one, two and three categorical variables against three, two and one numeric variables accordingly. This was produced for datasets containing six, where one to five

CC BY: Attribution alone 4.0
https://creativecommons.org/licenses/by/4.0/
https://doi.org/10.26219/heal.aueb.5832

categorical variables were included, and eight variables, where one to seven categorical variables were included.

The datasets containing four variables were tested on 54 datasets, six variables on 90 datasets, and eight variables on 126 datasets. These datasets include variation in the level of difficulty, between easy, medium and hard, while also a variation in minimal between-cluster distance, between 0.1, 0.3 and 0.5. The measure were implemented on each of these datasets, and the average of the Rand index for each combination was calculated.

The results for the datasets containing four variables are shown in *Table 4.8*, which is the average Rand index of 54 Rand indices, as they were calculated from a total of 54 datasets each. For six variables the average of 90 Rand indices is shown in *Table 4.10*. Similarly, the average of 126 Rand indices for eight variables is shown in *Table 4.12*. The variance of these indices is shown in *Table 4.9*, *Table 4.11* and *Table 4.13*, respectively.

# of var.	*#of cat var*	*Gower*	*G_IQ*	*G_ID*	*G_KDE*	*G_KNN*	*G_CD*
4	1	0.669	0.708	0.793	0.344	0.602	0.882
	2	0.703	0.658	0.715	0.392	0.613	0.379
	3	0.680	0.496	0.552	0.478	0.578	0.284
		G_IOF	*G_Lin*	*G_VE*	*G_Wishart*	*G_Podani*	*Laflin*
	1	0.628	0.625	0.616	0.775	0.632	0.636
	2	0.690	0.697	0.647	0.849	0.700	0.693
	3	0.676	0.586	0.602	0.642	0.680	0.680
		Huang	*Harikumar-PV*	*Ahmad & Dey*	*Modha & Sprangler*	*SBAC*	*TreeClust*
	1	0.901	0.856	0.856	0.842	0.606	0.792
	2	0.829	0.821	0.819	0.839	0.527	0.795
	3	0.616	0.615	0.608	0.606	0.520	0.801

Table 4.8 Average value of Rand index for four variables broken down by number of categorical variables.

CC BY: Attribution alone 4.0
https://creativecommons.org/licenses/by/4.0/
https://doi.org/10.26219/heal.aueb.5832

# of var.	#of cat var	Gower	G_IQ	G_ID	G_KDE	G_KNN	G_CD
4	1	0.006	0.025	0.033	0.007	0.003	0.010
	2	0.005	0.044	0.029	0.019	0.008	0.013
	3	0.010	0.040	0.032	0.042	0.021	0.02
		G_IOF	G_Lin	G_VE	G_Wishart	G_Podani	Laflin
	1	0.004	0.003	0.004	0.020	0.004	0.003
	2	0.007	0.012	0.012	0.011	0.005	0.004
	3	0.010	0.023	0.013	0.008	0.010	0.010
		Huang	Harikumar-PV	Ahmad & Dey	Modha & Sprangler	SBAC	TreeClust
	1	0.012	0.011	0.01	0.012	0.005	0.009
	2	0.011	0.017	0.015	0.010	0.019	0.010
	3	0.046	0.051	0.051	0.004	0.040	0.006

Table 4.9 Variance of Rand index for four variables broken down by number of categorical variables.

# of var	#of cat var	Gower	G_IQ	G_ID	G_KDE	G_KNN	G_CD
6	1	0.803	0.773	0.831	0.315	0.630	0.916
	2	0.814	0.726	0.732	0.404	0.644	0.420
	3	0.732	0.598	0.702	0.377	0.573	0.285
	4	0.740	0.641	0.720	0.434	0.565	0.295
	5	0.712	0.462	0.504	0.413	0.482	0.300
		G_IOF	G_Lin	G_VE	G_Wishart	G_Podani	Laflin
	1	0.704	0.695	0.695	0.854	0.782	0.699
	2	0.745	0.757	0.718	0.684	0.741	0.746
	3	0.727	0.697	0.696	0.714	0.675	0.666
	4	0.740	0.805	0.800	0.817	0.723	0.715
	5	0.703	0.655	0.662	0.744	0.718	0.712
		Huang	Harikumar-PV	Ahmad & Dey	Modha & Sprangler	SBAC	TreeClust
	1	0.869	0.840	0.833	0.853	0.653	0.786
	2	0.749	0.761	0.746	0.674	0.542	0.781
	3	0.826	0.848	0.816	0.667	0.549	0.789
	4	0.663	0.682	0.667	0.797	0.457	0.810
	5	0.783	0.814	0.797	0.688	0.393	0.794

Table 4.10. Average value of Rand index for six variables broken down by number of categorical variables.

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

CC BY: Attribution alone 4.0
https://creativecommons.org/licenses/by/4.0/
https://doi.org/10.26219/heal.aueb.5832

# of var	#of cat var	Gower	G_IQ	G_ID	G_KDE	G_KNN	G_CD
6	1	0.023	0.039	0.034	0.012	0.003	0.014
	2	0.012	0.031	0.028	0.019	0.007	0.010
	3	0.006	0.025	0.026	0.020	0.011	0.002
	4	0.010	0.039	0.028	0.028	0.013	0.003
	5	0.009	0.029	0.036	0.027	0.025	0.010
		G_IOF	G_Lin	G_VE	G_Wishart	G_Podani	Laflin
	1	0.013	0.011	0.011	0.016	0.027	0.012
	2	0.010	0.020	0.014	0.009	0.005	0.006
	3	0.006	0.025	0.025	0.003	0.009	0.008
	4	0.010	0.019	0.027	0.019	0.014	0.006
	5	0.008	0.024	0.032	0.009	0.009	0.009
		Huang	Harikumar-PV	Ahmad & Dey	Modha & Sprangler	SBAC	TreeClust
	1	0.013	0.018	0.015	0.018	0.017	0.018
	2	0.016	0.016	0.016	0.008	0.018	0.019
	3	0.03	0.04	0.013	0.006	0.038	0.012
	4	0.059	0.067	0.062	0.009	0.029	0.013
	5	0.06	0.08	0.065	0.01	0.095	0.04

Table 4.11 Variance of Rand index for six variables broken down by number of categorical variables.

CC BY: Attribution alone 4.0
https://creativecommons.org/licenses/by/4.0/ https://doi.org/10.26219/heal.aueb.5832

# of var	#of cat var	Gower	G_IQ	G_ID	G_KDE	G_KNN	G_CD
8	1	0.736	0.879	0.878	0.330	0.546	0.970
	2	0.735	0.780	0.807	0.342	0.667	0.530
	3	0.718	0.861	0.851	0.718	0.718	0.718
	4	0.718	0.718	0.718	0.327	0.612	0.271
	5	0.769	0.642	0.772	0.769	0.769	0.769
	6	0.756	0.756	0.756	0.357	0.505	0.276
	7	0.749	0.374	0.503	0.353	0.444	0.264
		G_IOF	G_Lin	G_VE	G_Wishart	G_Podani	Laflin
	1	0.612	0.599	0.600	0.793	0.603	0.608
	2	0.692	0.651	0.632	0.630	0.703	0.654
	3	0.718	0.718	0.718	0.718	0.718	0.662
	4	0.733	0.686	0.692	0.623	0.697	0.680
	5	0.769	0.769	0.769	0.769	0.769	0.719
	6	0.762	0.713	0.699	0.748	0.742	0.715
	7	0.755	0.535	0.505	0.596	0.720	0.749
		Huang	Harikumar-PV	Ahmad & Dey	Modha & Sprangler	SBAC	TreeClust
	1	0.982	0.874	0.873	0.803	0.550	0.795
	2	0.863	0.933	0.919	0.703	0.593	0.759
	3	0.818	0.632	0.642	0.661	0.631	0.668
	4	0.680	0.721	0.728	0.622	0.431	0.762
	5	0.756	0.719	0.719	0.719	0.719	0.719
	6	0.715	0.713	0.700	0.723	0.372	0.778
	7	0.574	0.553	0.544	0.564	0.755	0.635

Table 4.12 Average value of Rand index for eight variables broken down by number of categorical variables.

CC BY: Attribution alone 4.0
https://creativecommons.org/licenses/by/4.0/
https://doi.org/10.26219/heal.aueb.5832

# of var	#of cat var	Gower	G_IQ	G_ID	G_KDE	G_KNN	G_CD
8	1	0.025	0.015	0.020	0.09	0.003	0.008
	2	0.008	0.028	0.019	0.007	0.008	0.023
	3	0.01	0.023	0.018	0.010	0.010	0.010
	4	0.009	0.009	0.009	0.007	0.004	0.013
	5	0.008	0.039	0.028	0.008	0.008	0.01
	6	0.009	0.009	0.009	0.021	0.018	0.01
	7	0.011	0.016	0.040	0.012	0.026	0.03
		G_IOF	*G_Lin*	*G_VE*	*G_Wishart*	*G_Podani*	*Laflin*
	1	0.004	0.004	0.004	0.004	0.002	0.002
	2	0.003	0.020	0.019	0.032	0.0054	0.006
	3	0.010	0.010	0.010	0.010	0.010	0.002
	4	0.007	0.024	0.020	0.003	0.010	0.007
	5	0.008	0.008	0.008	0.008	0.008	0.008
	6	0.006	0.024	0.029	0.005	0.012	0.007
	7	0.017	0.045	0.028	0.007	0.006	0.011
		Huang	*Harikumar-PV*	*Ahmad & Dey*	*Modha & Sprangler*	*SBAC*	*TreeClust*
	1	0.001	0.030	0.023	0.006	0.001	0.02
	2	0.014	0.014	0.012	0.008	0.01	0.009
	3	0.050	0.002	0.002	0.002	0.002	0.006
	4	0.007	0.007	0.006	0.001	0.02	0.007
	5	0.006	0.008	0.008	0.008	0.008	0.008
	6	0.007	0.063	0.062	0.004	0.02	0.010
	7	0.016	0.012	0.026	0.008	0.01	0.045

Table 4.13 Variance of Rand index for eight variables broken down by number of categorical variables.

The majority of the measures performance reduces as the categorical variables of the dataset increase. For Gower's similarity, G_IOF, G_Lin, G_VE, G_Wishart, G_Podani and TreeClust measures, a small change in the Rand index is observed compared to the other measures.

The G_CD and Huang measures are seen to outperform the rest of the measures, when the datasets contain only one categorical variable. The Huang measure maintains its performance with an increase from one categorical variable to two and continues to have a high Rand index. The Harikumar-PV, Ahmad & Dey and Huang measures show high performance for

CC BY: Attribution alone 4.0
https://creativecommons.org/licenses/by/4.0/ https://doi.org/10.26219/heal.aueb.5832

all variations. A drop is observed for these in some cases, when all variables are categorical, apart from one numeric variable, for example from two categorical variables to three in the four variable datasets. In this case Gowers similarity and G_Podani maintain a high performance.

4.2.5 Conclusions

Based on the situations described in this paragraph, the results for Rand indices of all are presented in a condensed form in *Table 4.14*. From this table, where the average for all datasets is calculated, and it is shown that on average the Rand index prefers the Huang, Harikumar-PV, Ahmad & Dey and TreeClust measures, whilst the traditional Gower's similarity also has a high performance.

Gower	*G_IQ*	*G_ID*	*G_KDE*	*G_KNN*	*G_CD*
0.735	0.647	0.719	0.372	0.578	0.443
G_IOF	*G_Lin*	*G_VE*	*G_Wishart*	*G_Podani*	*Laflin*
0.707	0.678	0.646	0.728	0.704	0.686
Huang	*Harikumar-PV*	*Ahmad & Dey*	*Modha & Sprangler*	*SBAC*	*TreeClust*
0.786	0.785	0.775	0.719	0.495	0.786

Table 4.14 Average values for Rand index – Experiment 1.

From these conclusions and the previous testing done within this section, it is concluded that Huang, Harikumar-PV, Ahmad & Dey, Modha & Sprangler and Wishart measures have a similar performance, which outperforms the rest of the measures. Gower, G_IQ, G_ID, G_IOF, G_Lin, G_VE, G_Podani, Laflin, SBAC and TreeClust come in second. G_CD performance worsen with more categorical variables within the dataset, while G_KDE seems to have the worst score compared to all methods.

4.3 Second Experiment – High Number of Variables

A second experiment was carried out in order to test the measure presented in Chapter 3, when there is a higher amount of variables included in the dataset. To be able to achieve this, the tested simulated datasets are comprised by a total of 35 variables. Specifically, 15 continuous, 15 categorical and 5 binary variables were included.

The minimal between-cluster distance was set to 0.3, ensuring an average clarity of cut between clusters. This is accomplished in the R package by setting *sepVal* to the equivalent distance. The categorical variables had three to seven categories and the total number of

UNIVERSITY OF ECONOMICS & BUSINESS · ATHENS · ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

CC BY: Attribution alone 4.0
https://creativecommons.org/licenses/by/4.0/
https://doi.org/10.26219/heal.aueb.5832

objects in the produced datasets varied from 300 to 700. This dataset was reproduced twenty times, to guarantee the robustness of the acquired results.

The measures were then calculated for the twenty datasets, and the average Rand index for each measure was computed. *Table 4.15* shows the average values of the 20 Rand indices for each measure.

Gower	*G_IQ*	*G_ID*	*G_KDE*	*G_KNN*	*G_CD*
0.716	0.755	0.756	0.477	0.432	0.275
G_IOF	*G_Lin*	*G_VE*	*G_Wishart*	*G_Podani*	*Laflin*
0.701	0.786	0.762	0.633	0.696	0.706
Huang	*Harikumar-PV*	*Ahmad & Dey*	*Modha & Sprangler*	*SBAC*	*TreeClust*
0.744	0.818	0.729	0.619	0.552	0.745

Table 4.15 Average values for Rand index – Experiment 2

From the results it is shown that the Harikumar-PV measure has the highest performance compared to the rest. Leading from this measure, the G_Lin, G_VE, G_IQ, G_ID, Huang, Ahmad & Dey and the TreeClust measures, also show a high performance for clustering the datasets. The Gower, G_IOF, G_Podani, Laflin measures come in third based on their performance compared to the other measures, also with an acceptable performance as the Rand index is above 0.7. The G_Wishart and Modha & Sprangler measures have an average performance. A poor performance is observed for G_KDE, G_KNN and SBAC, while the worst performance is found for G_CD.

The variance of the Rand indices is shown in *Table 4.16*.

Gower	*G_IQ*	*G_ID*	*G_KDE*	*G_KNN*	*G_CD*
0.001	0.003	0.005	0.004	0.003	0.002
G_IOF	*G_Lin*	*G_VE*	*G_Wishart*	*G_Podani*	*Laflin*
0.002	0.004	0.008	0.007	0.003	0.002
Huang	*Harikumar-PV*	*Ahmad & Dey*	*Modha & Sprangler*	*SBAC*	*TreeClust*
0.020	0.029	0.023	0.007	0.004	0.010

Table 4.16 Variance of Rand indices – Experiment 2

CC BY: Attribution alone 4.0
https://creativecommons.org/licenses/by/4.0/ https://doi.org/10.26219/heal.aueb.5832

CHAPTER 5

Conclusions

This master's thesis focuses on the similarity measures involving clustering mixed data. A common strategy for clustering a data set with mixed variables is by first stating a distance measure for use in the clustering algorithm, with one of the most often used mixed data measures being Gower's similarity. The purpose of this thesis was to explore several modifications of Gower's similarity, as well as various other existing similarity measures proposed in the literature of mixed type variables. In Chapter 2, a brief review of the cluster analysis background was conducted. It was seen that although many clustering algorithms exist, solutions for mixed-type data are sparse. In Chapter 3, several existing distance measures for mixed data were presented. Ten modifications of Gower's coefficient were examined, as well as seven other proposed similarity measures.

The inability to verify a method against a known "truth" is a basic issue in the evaluation of clustering algorithms. Therefore, in Chapter 4, studies on simulated datasets were performed in order to test the evaluated measures. The produced datasets consisted of simulated data with four clear clusters, that were pre-defined. In a first experiment, 270 mixed data simulated datasets were including for the evaluation of the measures, comprised by various features falling into three different categories. Firstly, their reliance on the dataset difficulty, as defined by the number of categories of the categorical variables, was examined. Secondly, the similarity measures performance was studied for different minimal between-cluster distances. Lastly, the dependency on various combinations of quantitative and categorical characteristics was investigated. In a second experiment, a larger number of variables was included on a total of 20 datasets, while all had the same characteristics.

The similarity measures were applied by using the agglomerative hierarchical clustering algorithm (complete linkage) in order to produce the clustering solution. For all the assessments and comparisons, the average Rand Indices of the simulations containing different characteristics were computed. This index gives a proportion of pairs of objects that have been correctly clustered in the same group or correctly clustered into different groups.

The results showed that Gower's similarity performs well, especially in the case of high minimal between-cluster distance, while it doesn't seem to be impacted substantially by the

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

CC BY: Attribution alone 4.0
https://creativecommons.org/licenses/by/4.0/ https://doi.org/10.26219/heal.aueb.5832

dataset difficulty or variable combination. Thus, this measure can be recommended for datasets with a high number of categorical variables as well as a low number.

G_ID is affected by the difficulty, the minimal between-cluster distance and the variable combination of the dataset as its performance worsens as the dataset gets harder and as the categorical variables increase but improves as the minimal between-cluster distance increases. It's performance remains at good levels, though, for high number of total variables, as shown in the second experiment. This is also the case for G_IQ and G_KNN, but G_ID seems to outperform both in most settings.

Huang, Harikumar-PV, Ahmad & Dey, Modha & Sprangler follow the same pattern, with Harikumar-PV and Ahmad & Dey have high performance for most settings and top performance compared to the remaining measures in various datasets. For the second experiment, in which a higher number of variables are included in the datasets, the Harikumar-PV measure seems to clearly outperform the remaining measures, as the difference between their Rand indices increases. The Ahmad & Dey measure continuous to perform well for these datasets. Therefore, these two measures seem to have a good performance in general applications and can be recommended for general use. The Modha & Sprangler measure seems to have a lower performance when the number of variables included in the dataset is high.

G_IOF is affected moderately by the dataset difficulty and minimal between-cluster distance, while its performance also slightly increases with additional variables but remains stable as the number of categorical variables within the dataset increases. Its overall performance is good for both experiments. G_Lin and G_VE are affected by the difficulty and minimal between-cluster distance but are considered stable for a change in the total number of variables and the number of categorical variables within the dataset. Their overall performance is moderate.

G_Podani, Laflin and TreeClust also react as the previous measures to change in difficulty, minimal between-cluster distance and variables, but it is shown that they have an overall good performance to the different datasets. Therefore, these measures can also be suggested for datasets with a high number of categorical variables. This point is also observed within the second experiment, as these measure show a good performance even when the total of variables is thirty-five.

UNIVERSITY OF ECONOMICS & BUSINESS ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ ATHENS

CC BY: Attribution alone 4.0
https://creativecommons.org/licenses/by/4.0/ https://doi.org/10.26219/heal.aueb.5832

G_Wishart is unaffected by the dataset difficulty level, has a higher performance for higher minimal between-cluster distance while its performance worsens with the increase of categorical variables within the dataset. Its overall performance is good within the first experiment but suffers from a drop when higher number of variables are included. Lastly, the G_KDE, G_CD, SBAC measures cannot be recommended for a regular use in hierarchical clustering because their cluster quality show poor performance.

Overall, in the first experiment the Harikumar-PV and Ahmad & Dey measures seem to be in the lead compared to the other measures, but with a slight difference in their scores. The difference between the Harikumar-PV and the remaining measures is clearer within the second experiment, for a higher number of variables. Therefore, Harikumar-PV and Ahmad & Dey are shown to outperform the remaining measures in most circumstances.

In the future, research on mixed data clustering is expected to continue to be driven by the development of algorithms to handle mixed data. As an option for research and development of the methods explored in this work, it is also suggested that different weight functions be applied.

CC BY: Attribution alone 4.0
https://creativecommons.org/licenses/by/4.0/
https://doi.org/10.26219/heal.aueb.5832

CHAPTER 6

References

Ackermann, H. (1997). A note on circular nonparametrical classification, *Biometrical journal*, 39(5), 577-587

Aggarwal, C. C. (2018). An introduction to cluster analysis. In Data Clustering, *Chapman and Hall/CRC*, 1-28

Ahmad, A., & Dey, L. (2007). A k-mean clustering algorithm for mixed numeric and categorical data, *Data & Knowledge Engineering*, 63(2), 503-527

Ahmad, A., & Khan, S. S. (2019). Survey of state-of-the-art mixed data clustering algorithms, *Ieee Access*, 7, 31883-31902

Azam, N. F., & Viktor, H. L. (2011). Spectral clustering: An explorative study of proximity measures, International Joint Conference on Knowledge Discovery, Knowledge Engineering, and Knowledge Management. *Springer, Berlin, Heidelberg*, 60-78.

Balaji, K., & Lavanya, K. (2018). Clustering algorithms for mixed datasets: a review, *International Journal of Pure and Applied Mathematics*, 18(7), 547-56

Batra, A. (2013). Analysis and approach: K-means and K-medoids data mining algorithms, *ICACCT, 5th IEEE International Conference on Advanced Computing & Communication Technologies*

Berkhin, P. (2006). A survey of clustering data mining techniques. In Grouping multidimensional data, *Springer, Berlin, Heidelberg*, 25-71

Berry, M. W., Mohamed, A., & Yap, B. W. (2019). Supervised and unsupervised learning for data science, *Springer Nature*

Bishnoi, S., & Hooda, B. K. (2020). A survey of distance measures for mixed variables

Brea, M. E. A. (2013). Constrained Clustering Algorithms: Practical Issues and Applications: PHD Thesis, *Universidade da Coruña, Departamento de Computación*

Breiman, L., J. H. Friedman, RA Olshen, and CJ Stone (1984). Classification and Regression trees, 23-32

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

CC BY: Attribution alone 4.0
https://creativecommons.org/licenses/by/4.0/
https://doi.org/10.26219/heal.aueb.5832

Budiaji, W. (2019). kmed: Distance-Based k-Medoids. R package version (2). https://cran.r-project.org/web/packages/kmed/kmed.pdf

Buttrey, S. E. (2006). "A scale-independent clustering method with automatic variable selection based on trees." *Joint Statistical Meetings, Seattle, WA*

Buttrey, S. E., & Whitaker, L. R. (2015). treeClust: An R Package for Tree-Based Clustering Dissimilarities. R Journal, 7(2).
https://cran.r-project.org/web/packages/treeClust/treeClust.pdf

Buttrey, S. E., & Whitaker, L. R. (2016). A Scale Independent, Noise Resistant Dissimilarity for Tree Based Clustering of Mixed Data. *Naval Postgraduate School Monterey United States*

Celebi, M. E., & Kingravi, H. A. (2015). Linear, deterministic, and order-invariant initialization methods for the k-means clustering algorithm. In Partitional clustering algorithms. *Springer, Cham*

Chae, S. S., Kim, J. M., & Yang, W. Y. (2006). Cluster analysis with balancing weight on mixed-type data. *Communications for Statistical Applications and Methods*, 13(3), 719-732

Cheung, Y. M., & Jia, H. (2013). Categorical-and-numerical-attribute data clustering based on a unified similarity metric without knowing cluster number. *Pattern Recognition*, 46(8), 2228-2238

Chiodi, M. (1990). A partition type method for clustering mixed data. *Rivista di statistica applicata*, 2, 135-147

Cord, M., & Cunningham, P. (2008). Machine learning techniques for multimedia: case studies on organization and retrieval. *Springer Science & Business Media*

Day, W. H., & Edelsbrunner, H. (1984). Efficient algorithms for agglomerative hierarchical clustering methods. *Journal of classification*, 1(1), 7-24

Depa, P., & Nivedita, N. (2013). "Pattern Recognition Algorithms for Cluster Identification Problem". *International Journal of Computer Science & Informatics (IJCSI)*

Dinh, D. T., Huynh, V. N., & Sriboonchitta, S. (2021). Clustering mixed numerical and categorical data with missing values. *Information Sciences*, 571, 418-442

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

CC BY: Attribution alone 4.0
https://creativecommons.org/licenses/by/4.0/
https://doi.org/10.26219/heal.aueb.5832

Dinler, D., & Tural, M. K. (2016). A survey of constrained clustering. In Unsupervised learning algorithms. *Springer, Cham*, 207-235

D'Orazio, M. (2021). Distances with mixed type variables some modified Gower's coefficients

D'Orazio, M., & D'Orazio, M. M. (2022). Package 'StatMatch'. https://cran.r-project.org/web/packages/StatMatch/StatMatch.pdf

Duda, R. O., Hart, P. E. & Stork D.G. (2001). Pattern classification. New York John Wiley & Sons

Everitt Brian (1980). Cluster Analysis. *Wiley*

Fan, J., Han, F., & Liu, H. (2014). Challenges of big data analysis. *National science review*, 1(2), 293-314

Fayyad, U. M., Piatetsky-Shapiro, G., & Smyth, P. (1996). Knowledge Discovery and Data Mining: Towards a Unifying Framework. *KDD*, 96, 82-88

Foss, A. H., & Markatou, M. (2018). kamila: clustering mixed-type data in R and Hadoop. *Journal of Statistical Software*, 83, 1-44
https://cran.r-project.org/web/packages/kamila/kamila.pdf

Foss, A. H., Markatou, M., & Ray, B. (2019). Distance metrics and clustering methods for mixed-type data. *International Statistical Review*, 87(1), 80-109

Foss, A., Markatou, M., Ray, B., & Heching, A. (2016). A semiparametric method for clustering mixed data. *Machine Learning*, 105(3), 419-458

Friedman, J. H., & Meulman, J. J. (2004). Clustering objects on subsets of attributes (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(4), 815-849

Gan, G., Ma, C., & Wu, J. (2007). Data clustering - theory, algorithms, and applications

Gan, G., Ma, C., & Wu, J. (2020). Data clustering: theory, algorithms, and applications. *Society for Industrial and Applied Mathematics*

Goodall, D. W. (1966). A new similarity index based on probability. *Biometrics*, 882-907

CC BY: Attribution alone 4.0
https://creativecommons.org/licenses/by/4.0/ https://doi.org/10.26219/heal.aueb.5832

Gordon, A. D. (1999). Classification. *CRC Press*

Gower, J. C. (1971). A general coefficient of similarity and some of its properties. *Biometrics*, 857-871

Gramacy, R. B. (2016). laGP: large-scale spatial modeling via local approximate Gaussian processes in R. *Journal of Statistical Software*, 72, 1-46.
https://cran.r-project.org/web/packages/laGP/laGP.pdf

Hahsler, M., Buchta, C., Gruen, B., & Hornik, K. (2022). Package 'arules'
https://cran.r-project.org/web/packages/arules/arules.pdf

Han, J. (2005). Data Mining: Concepts and Techniques. *San Francisco, CA, USA: Morgan Kaufmann Publishers Inc*

Han, J., Kamber, M. (2006). Data Mining: Concepts and Techniques. *2nd Edition, Morgan Kaufmann Publishers, San Francisco*

Han, J., Pei, J., & Kamber, M. (2011). Data mining: concepts and techniques. *Elsevier*

Handl, J., Knowles, J. & Kell, D. B. (2005). Computational cluster validation in post-genomic data analysis. *Bioinformatics*, 21(15), 3201–3212

Harikumar, S., &P.V., S. (2015). K-medoid clustering for heterogeneous datasets. *Procedia Computer Science*, 70, 226-237

Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). The elements of statistical learning: data mining, inference, and prediction. *New York: springer*, 2, 1-758

Heinz, G., Peterson, L. J., Johnson, R. W., & Kerk, C. J. (2003). Exploring relationships in body dimensions. *Journal of Statistics Education*, 11(2)

Hendrickson, J. (2014). Methods for clustering mixed data. *Doctoral dissertation, University of South Carolina*

Hennig, C., & Liao, T. F. (2013). How to find an appropriate clustering for mixed-type variables with application to socio-economic stratification. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 62(3), 309-369

Hsu, C. C., Chen, C. L., & Su, Y. W. (2007). Hierarchical clustering of mixed data based on distance hierarchy. *Information Sciences*, 177(20), 4474-4492

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

CC BY: Attribution alone 4.0
https://creativecommons.org/licenses/by/4.0/
https://doi.org/10.26219/heal.aueb.5832

Huang, Z. (1997). Clustering large data sets with mixed numeric and categorical values. In Proceedings of the 1st pacific-asia conference on knowledge discovery and data mining. *PAKDD*, 21-34

Huang, Z. (1998). Extensions to the k-means algorithm for clustering large data sets with categorical values. *Data mining and knowledge discovery*, 2(3), 283-304

Hunt, L., & Jorgensen, M. (2011). Clustering mixed data. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1(4), 352-361

Jain, A. K., & Dubes, R. C. (1988). Algorithms for clustering data. *Prentice-Hall, Inc.*

Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: a review. *ACM computing surveys (CSUR)*, 31(3), 264-323

Jing, R., Zhao, X., & Yu, J. (2007). A Theoretical Analysis Framework of Customer Cluster Value of Dynamic B2C E-Commerce. *International Conference on Wireless Communications, Networking and Mobile Computing, IEEE*, 3438-3441

Kasif, S., Salzberg, S., Waltz, D., Rachlin, J., & Aha, D. W. (1998). A probabilistic framework for memory-based reasoning. *Artificial Intelligence*, 104(1-2), 287-311

Kassambara, A., & Mundt, F. (2017). Package 'factoextra'. Extract and visualize the results of multivariate data analyses, 76(2).
https://cran.r-project.org/web/packages/factoextra/factoextra.pdf

Kaufman, L. and Rousseeuw, P. (1990). Finding Groups in Data: An Introduction to Cluster Analysis. *New York, Wiley*

Kaufman, L., & Rousseeuw, P. J. (2009). Finding groups in data: an introduction to cluster analysis. *John Wiley & Sons*

King, B. (1967). Step-wise clustering procedures. *Journal of the American Statistical Association*, 62(317), 86-101

Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *The annals of mathematical statistics*, 22(1), 79-86

Laflin, S. (1998). Laflin's general coefficient

Larose, D. T., & Larose, C. D. (2004). Decision trees. Discovering knowledge in data: An introduction to data mining. *Wiley-Interscience*

CC BY: Attribution alone 4.0
https://creativecommons.org/licenses/by/4.0/
https://doi.org/10.26219/heal.aueb.5832

Li, C., & Biswas, G. (1998). Conceptual Clustering With Numeric-and-Nominal Mixed Data-A New Similarity Based System. *IEEE Transcript on KCE*

Li, C., & Biswas, G. (2002). Unsupervised learning with mixed numeric and nominal data. *IEEE Transactions on knowledge and data engineering*, 14(4), 673-690

Li, J., Ray, S. & Lindsay, B. (2007). A nonparametric statistical approach to clustering via mode identification. *J.Mach.Learn.Res*., 8, 1687–1723

Liao, S., Ho, H., Lin, H. (2008). Mining stock category association and cluster on Taiwan stock market. *Expert Systems with Applications*, 35, 19–29

Lin, D. (1998). An information-theoretic definition of similarity. *Proceedings of the 15th International Conference on Machine Learning, San Francisco: Morgan Kaufmann*, 296–304

MacQueen J. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, 281–297

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., Hornik, K., & Studer, M. (2013). Package 'cluster'. https://cran.r-project.org/web/packages/cluster/cluster.pdf

Mateo, R., Salvo, M. et al. (2008). Balanced Clustering using Mobile Agents for the Ubiquitous Healthcare Systems. *Proceedings of the 3rd International Conference on Convergence and Hybrid Information Technology, Busan, South Korea*, 686–691

Meila, M., & Shi, J. (2000). Learning segmentation by random walks. Advances in neural information processing systems

Milligan, G. W., & Cooper, M. C. (1985). An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, 50(2), 159-179

Modha, D. S., & Spangler, W. S. (2003). Feature weighting in k-means clustering. *Machine learning*, 52(3), 217-237

Mollaian, M., Dörgő, G., & Palazoglu, A. (2022). Performing Multi-Objective Optimization Alongside Dimension Reduction to Determine Number of Clusters. Processes, 10(5), 893

Montanari, A., & Mignani, S. (1994). Notes on the bias of dissimilarity indices for incomplete data sets: The case of archaelogical classification. 39-49

CC BY: Attribution alone 4.0
https://creativecommons.org/licenses/by/4.0/
https://doi.org/10.26219/heal.aueb.5832

Nocedal, J., & Wright, S. J. (1999). Numerical optimization. *New York, NY: Springer New York*

Petchey, O. L., & Gaston, K. J. (2009). Dendrograms and measures of functional diversity: a second instalment. *Oikos*, 118(7), 1118-1120

Pham, D. T., & Afify, A. A. (2006). Engineering applications of clustering techniques. *Intelligent Production Machines and Systems*, 326-331

Podani, J. (1999). Extending Gower's general coefficient of similarity to ordinal characters. *Taxon*, 48(2), 331-340

Qiu, W., & Joe, H. (2006). Separation index and partial membership for clustering. *Computational statistics & data analysis*, 50(3), 585-603

Qiu, W., & Joe, H. (2015). clusterGeneration: Random cluster generation (with specified degree of separation). R package version 1.3.4. https://cran.r-project.org/web/packages/clusterGeneration/clusterGeneration.pdf

Rafsanjani, K. M., Varzaneh, Z. A., & Chukanlo, N. E. (2012). A Survey Of Hierarchical *Clustering Algorithms. J. Math. Comput. Sci.*, 5(3), 229–240

Raftery, A. E., & Dean, N. (2006). Variable selection for model-based clustering. *Journal of the American Statistical Association*, 101(473), 168-178

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, 66(336), 846-850

Ren, M., Liu, P., Wang, Z., & Pan, X. (2016). An improved mixed-type data based kernel clustering algorithm. *12th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery*, 1205-1209

Reynolds, A. P., Richards, G., de la Iglesia, B., & Rayward-Smith, V. J. (2006). Clustering rules: a comparison of partitioning and hierarchical clustering algorithms. *Journal of Mathematical Modelling and Algorithms*, 5(4), 475-504

Rocha, M. R., Gaedke, U., & Vasseur, D. A. (2011). Functionally similar species have similar dynamics. *Journal of Ecology*, 99(6), 1453-1459

Silverman, BW (1986). Density Estimation for Statistics and Data Analysis. *Chapman & Hall/CRC, London*

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

CC BY: Attribution alone 4.0
https://creativecommons.org/licenses/by/4.0/
https://doi.org/10.26219/heal.aueb.5832

Sneath, P.H. and Sokal, R.R. (1973). Numerical Taxonomy. *London, UK: Freeman*

Sparck-Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1), 11–21

Šulc, Z. and Rezanková, H. (2015). Novel similarity measures for categorical data based on mutability and entropy. *Conference of the International Federation of Classification Societies*

Šulc, Z, Matejka, M., & Procházka, J. (2016). Modifications of the Gower similarity coefficient. *19th Conference of Applications of Mathematics and Statistics in Economics*

Šulc, Z., & Řezanková, H. (2019). Comparison of similarity measures for categorical data in hierarchical clustering. *Journal of Classification*, 36(1), 58-72

Sulc, Z., Cibulkova, J., & Rezankova, H. (2022). Nomclust 2.0: an R package for hierarchical clustering of objects characterized by nominal variables. *Computational Statistics*, 1-24.
https://cran.r-project.org/web/packages/nomclust/nomclust.pdf

Šulc, Z., Matějka, M., Procházka, J., & Řezanková, H. (2017). Evaluation of the Gower coefficient modifications in hierarchical clustering. *Advances in Methodology and Statistics*, 14(1), 37-48

Talburt, J. R. (2011). Entity resolution and information quality. *Elsevier*

Tan, P.-N., Steinbach, M., and Kumar, V. (2005). Introduction to Data Mining. *USA: Addison-Wesley Longman, Inc*

van de Velden, M., Iodice D'Enza, A., & Markos, A. (2019). Distance-based clustering of mixed data. *Wiley Interdisciplinary Reviews: Computational Statistics*, 11(3)

van den Hoven, J. (2015). Analyzing Spotify data: Exploring the possibilities of user data from a scientific and business perspective (Research Thesis). *Amsterdam: VU Amsterdam*

van den Hoven, J. (2015). Clustering with optimised weights for Gower's metric. *University of Amsterdam*

Vavrek, M. J., & Vavrek, M. M. J. (2020). Package 'fossil'.
https://cran.r-project.org/web/packages/fossil/fossil.pdf

CC BY: Attribution alone 4.0 https://doi.org/10.26219/heal.aueb.5832
https://creativecommons.org/licenses/by/4.0/

Vinh, N. X., Epps, J., & Bailey, J. (2009). Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *The Journal of Machine Learning Research*

Wang, S., Yabes, J. G., & Chang, C. C. H. (2021). Hybrid Density-and Partition-Based Clustering Algorithm for Data With Mixed-Type Variables. *Journal of Data Science*, 19(1), 15-36

Webb, A. R. (2002). Statistical pattern recognition. *John Wiley & Sons*

Wilson, DR., Martinez, TR. (1996). Heterogeneous radial basis functions. *In Proceedings of the International Conference on Neural Networks*

Wilson, DR., Martinez, TR. (1997). Improved heterogeneous distance functions. *Journal of Artificial Intelligence Research*

Wishart, D. (2003). K-means clustering with outlier detection, mixed variables and missing values. In Exploratory Data Analysis in Empirical Research. *Springer, Berlin, Heidelberg*, 216-226

Witten, I. H., & Frank, E. (2005). Data mining: practical machine learning tools and techniques with Java implementations. *Morgan Kaufmann*

Xu R, Wunsch D (2005). Survey of clustering algorithms. *IEEE Transactions on Neural Networks*, 16(3), 645–678

Xu, R., & Wunsch, D. (2008). Clustering. *John Wiley & Sons*

Yang, R. (2012). A hierarchical clustering and validity index for mixed data. *Graduate theses and dissertations*, 12534, P6

Zhang, W., & Di, Y. (2020). Model-based clustering with measurement or estimation errors. *Genes*, 11(2), 185

Zhao, Y., & Karypis, G. (2002). Evaluation of hierarchical clustering algorithms for document datasets. *In Proceedings of the eleventh international conference on Information and knowledge management*, 515-524

Zhu, X., & Goldberg, A. B. (2009). Introduction to semi-supervised learning. *Synthesis lectures on artificial intelligence and machine learning*, 3(1), 1-130

CC BY: Attribution alone 4.0
https://creativecommons.org/licenses/by/4.0/
https://doi.org/10.26219/heal.aueb.5832