A Framework of Unsupervised Machine Learning Algorithms for User Profiling

Erik Kuiper
University of Twente
Efthymios Constantinides
University of Twente
Sjoerd de Vries
University of Twente
Robert Marinescu-Muster
University of Twente
Floris Metzner
University of Twente

Cite as:
Kuiper Erik, Constantinides Efthymios, de Vries Sjoerd, Marinescu-Muster Robert, Metzner Floris (2019), A Framework of Unsupervised Machine Learning Algorithms for User Profiling. *Proceedings of the European Marketing Academy*, 48th, (10235)

Paper presented at the 48th Annual EMAC Conference, Hamburg, May 24-27, 2019.

A Framework of Unsupervised Machine Learning Algorithms for User Profiling

Abstract

Organizations often have difficulties to extract knowledge from data and selecting appropriate Machine Learning algorithms in order to develop accurate Behavioural Profiles or user segments. Moreover, marketing departments often lack a fundamental understanding on data-driven segmentation methodologies. This paper aims to develop a framework outlining Unsupervised Machine Learning algorithms for the purpose of User Profiling with respect to important data properties. A systematic literature review was conducted on the most prominent Unsupervised Machine Learning algorithms and their requirements regarding the characteristics of the dataset.

A framework is proposed outlining various Unsupervised Machine Learning algorithms for User Profiling. It provides two-stage clustering strategies for categorical, numerical, and mixed types of data with respect to the data size and data dimensionality. The first stage consists of an hierarchical or model-based clustering algorithm to determine the number of clusters. In the second stage, a non-hierarchical clustering algorithm is applied for cluster refinement.

The framework can support researchers and practitioners to determine which Unsupervised Machine Learning algorithms are appropriate for developing robust behavioural profiles or data-driven user segments.

Keywords: Unsupervised Machine Learning – Data-Driven Segmentation – Digital Marketing

Track: Digital Marketing & Social Media

1. INTRODUCTION

Advancements in the Internet of Things, Neuroscience, Artificial Intelligence, and Data Mining have propelled the desire and collection of data for strategic decision making and personalization. A key competitive advantage for today's organizations is the availability of large amounts of data for the purpose of segmenting a customer base, offering tailored services, and extracting meaningful information provided by various data sources (Chester, 2012). Machine Learning (ML) plays a key role in data mining applications to gain insights from unstructured data. According to Bose and Mahapatra (2001) ML is "the study of computational methods to automate the process of knowledge acquisition from examples" (p. 212). ML can be divided into *unsupervised* and *supervised* machine learning (Larose, 2014; Prasad, 2016).
In *unsupervised Machine Learning* (UML), no target variable is specified and only input data are provided (Larose, 2014). In contrast, *Supervised Machine Learning* (SML) algorithms are given a specific goal (e.g., target variable) for grouping data (Larose, 2014; Prasad, 2016; Walter & Bekker, 2017). This paper focuses on UML for data-driven customer segmentation and user profiling. *User Profiling* can be referred to as the process of gathering information specific to each individual either explicitly or implicitly (Eirinaki &Vazirgiannis, 2003). A user profile generally includes geo-demographic, psychographic, or behavioural information (Eirinaki &Vazirgiannis, 2003).

However, organizations are often unable to gain meaningful insights out of data whereby a considerable amount of opportunities, resources, and marketing efforts are wasted. In addition, marketing departments often lack a fundamental understanding on data-driven segmentation methodologies (e.g., Dolnicar, 2009; Boratto et al., 2016). Key issues in methodological decisions for data-driven segmentation are determining the number of clusters and which clustering algorithm should be chosen (Dolnicar, 2009). In addition, approaches for numerical data are relatively well-understood and widely available but approaches for categorical or mixed types of data are less prevalent and straightforward (Boriah, 2008). For instance, in contrast to numerical data, categorical data is deficient of default ordering relationships on the attribute values which make the task of developing distance measures and clustering algorithms for categorical data more challenging (Alamuri, Surampudi, & Negi, 2014). Prior research focused on the development, effectiveness (i.e., accuracy), and efficiency of various UML algorithms (e.g., Tamasauskas et al., 2012; Pandove et al., 2018, Park et al., 2009). Moreover, most clustering algorithms can either handle large data sets but are limited to only handling numerical or categorical attributes or they are able to handle both types of data but are inefficient at handling large datasets (Fahad et al., 2014). Selecting an appropriate algorithm is therefore a difficult and time consuming task. Important aspects to consider are the research question being addressed, variables used to characterize objects, the data type, data size, data dimensionality, distance measures, and outliers (Han, Kamber, & Pei, 2012; Larose, 2014). However, none of the studies provided an outline of UML algorithms and the various requirements and characteristics regarding the dataset.

This paper aims to contribute towards an answer by developing a methodology and framework of UML algorithms, based on a two-stage clustering methodology, with respect to important data properties. The framework is aimed at supporting researchers and practitioners in selecting the most appropriate algorithms and consequently obtaining accurate segmentation results. The research question is as follows: *What is an appropriate framework for outlining Unsupervised Machine Learning Algorithms for User Profiling?*

2. METHODOLOGY

A systematic literature review will be conducted according to the methods described in Wolfswinkel, Furtmueller and Wilderom (2013) and Webster and Watson (2002). Reviewing the core concepts of UML algorithms enables the researcher to develop a methodology and a framework. Different scientific search engines are considered including Scopus, Web of

Science, and Google Scholar. Articles are filtered by relevance and a first selection is done by evaluating the title, abstract, and publication date. Next, the articles are compared by the amount of citations and finally by reading the full text. The literature review is organized thematically. The literature review is provided in chapter 3 and the *framework* in subsection 3.6. In chapter 4, the discussion, theoretical and practical implications, future research, and research limitations are provided.

3. LITERATURE REVIEW

3.1 Machine Learning

The beginning of artificial intelligence (AI) in academic literature can be found around 1950 wherein Turing (1950) wrote the paper Computing Machinery and Intelligence. Within AI, Machine Learning (ML) has become the technology of choice in achieving practical solutions (Jordan & Mitchell, 2015). They argue that the fast decrease in the cost of computational power and the availability of accumulating amounts of data are the two factors that drive the developments in ML. ML can play a key role in data mining applications to gain insights from unstructured data. According to Bose and Mahapatra (2001) ML is "the study of computational methods to automate the process of knowledge acquisition from examples" (p. 212). An important feature is that ML is not programmed to follow particular decision rules to create results, but rather, it has the capability of creating those rules by data and feedback (Jordan & Mitchell, 2015). ML techniques can be divided into two main categories of *unsupervised* and *supervised* learning, which are reviewed the following sections (Larose, 2014; Prasad, 2016).

3.1.1 Unsupervised Machine Learning

In Unsupervised Machine Learning, no target variable is specified and only input data are provided (Larose, 2014). *Clustering* and its variations are often referred to as Unsupervised Machine Learning (Larose, 2014; Prasad, 2016; Walter & Bekker, 2017). Clustering is a multivariate technique whose primary purpose is to group objects so that each object is similar to the other objects in the cluster and different from objects in all the other clusters (Larose, 2014; Prasad, 2016). Examples are understanding consumer behaviour by identifying homogeneous groups of customers, identifying new product opportunities by clustering products or brands, relationship identification, or for data reduction purposes. Clustering can be regarded as market segmentation which is one of the most central strategic issues in marketing (Dolnicar, 2002). The success of targeted marketing activities depend on the quality of the (data-driven) market segments constructed. Hence, a benefit of clustering lies in being able to tailor an organisations offerings with the needs of a particular customer group, in doing so, the organization gains a competitive advantage in the marketplace (Dolnicar 2008; Hiziroglu 2013). Important issues and requirements for clustering analysis are the research question being addressed, variables used to characterize objects, data type, data size, data dimensionality, distance measures, outlier detection, and the interpretability (Han, Kamber, & Pei, 2012; Larose, 2014).

The major fundamental clustering algorithms can be classified as: (1) Hierarchical-based, (2) Partitioning-based, (3) Density-based, (4) Grid-based, and (5) Model-based (Han et al., 2012; Fahad et al., 2014). In Density-based methods objects are separated based on their density, connectivity, and boundary (Fahad et al., 2014). Here, the density of objects is analysed to determine the functions of datasets that influence a particular object. In Grid-based methods the space of the data objects are separated into grids. In Model-based methods the fit between the data and a predefined mathematical model is optimized based on the assumption that the data includes a mixture of underling probability distributions (Fahad et al., 2014; Han et al., 2012). Model-based methods are able to automatically determine the number of clusters and taking outliers into account. Examples are Neural Networks such as Self-Organising Maps developed by Kohonen (1982).

This paper is limited to reviewing *Hierarchical-based* and *Partitioning-based* methods. Moreover, Dolnicar (2002) studied the standards of various clustering methods used in academic literature and found that the majority of segmentation applications (73%) either used hierarchical or non-hierarchical (i.e., partitioning) methods.

3.1.2 Hierarchical and Non-Hierarchical Clustering

Hierarchical clustering methods are aimed at finding a structure in the data (i.e., a hierarchy) depending on the medium of proximity and are represented in a tree-like structure known as a dendrogram. Hierarchical clustering can be either agglomerative (i.e., bottom-up) or divisive (i.e., top-down). Agglomerative clustering initiates with one object for each cluster and reclusively merges it with two or more similar clusters (Fahad, 2014). A divisive variant operates in the opposite direction, wherein it initiates with the dataset as one cluster and reclusively separates objects to the most appropriate clusters (Fahad, 2014). However, drawbacks of hierarchical methods are that they cannot handle large datasets or high dimensionality well (Fahad, 2014; Pandove, Goel, & Rani, 2018). An advantage of hierarchical methods is that it is not required to specify the number of clusters a-priori. Furthermore, five agglomerative approaches exist including Single Linkage, Complete Linkage, Average Linkage, Centroid's method, and Ward's method (Fahad et al., 2014; Tamasauskas et al., 2012).

Non-hierarchical clustering algorithms divide data objects into several partitions where each partition represents a cluster. Non-hierarchical methods are commonly used for handling large datasets because they are computationally less expensive (Fahad et al., 2014; Pandove, 2018). Non-hierarchical clustering can be *Hard* or *Soft* (Prasad, 2016). The basic methods typically adopt hard clustering known as *exclusive cluster separation* (Han et al., 2012). Here, each object must belong to exactly one group. In soft methods this requirement is relaxed by techniques such as fuzzy clustering.

3.1.4 Unsupervised Machine Learning Algorithms

Determining the algorithm and similarity measure to calculate the distance between objects is a key step for clustering analysis. Similarity measures for continuous data are relativity well-understood and widely available but for categorical data it is not as straight forward (Boriah, 2008). In contrast to continuous data, categorical data is deficient of default ordering relationships on the attribute values which make the task of developing distance measures and clustering algorithms for categorical data more challenging (Alamuri, Surampudi, & Negi, 2014). A distinctive characteristic of data mining applications is that it deals with large, complex, or high dimensional datasets. Datasets can include millions of objects and hundreds of attributes. Hence, ML algorithms are therefore required to be scalable and capable of handling different types of attributes. Interesting clustering algorithms are those who can handle large datasets of numeric or categorical variables because these types of data are most frequently present in real world data (Dolnicar, 2002). However, most clustering algorithms can either handle large data sets but are limited to numeric attributes or they are able to handle both types of data but are inefficient at handling large datasets (Fahad et al., 2014).

For *non-hierarchical clustering*, MacQueen (1967) introduced the *k-means algorithm* which can efficiently handle large datasets and is therefore well suited for data mining tasks. In the k-means algorithm the centre is the average of all points representing the arithmetic mean (Fahad et al., 2014). It iteratively searches the cluster centres and updates the memberships of objects to minimise the within cluster sum of squares (WCSS) using the *(squared) Euclidean distance* measure. A drawback is that k-means only works efficiently on *numerical* data (MacQueen, 1967; Fahad et al., 2014). Huang (1998) introduced the *k-modes* non-hierarchical algorithm which is suitable for clustering large *categorical* datasets. The key differences are that k-modes uses a *simple matching dissimilarity measure* (i.e., hamming distance) instead of Euclidean distance, replaces the means of clusters with modes, and uses a frequency-based method to update cluster modes (Huang, 1998). The k-modes dissimilarity measure is defined by the total

mismatches of corresponding attribute categories of the two objects (Huang, 1998). Hence, the smaller the amount of mismatches the higher the similarity between objects. Furthermore, k-modes is faster compared to k-means because it converges in less iterations (Huang, 1998). A similar algorithm is *k-medoids* introduced by Park and Jun (2009) wherein medoids are considered instead of centroids or modes. It is based on the most centrally located object within a cluster and therefore less sensitive to outliers (Park & Jun, 2009). Hence, k-medoids is suitable for *categorical* data and handling outliers (i.e., noise) but it does not handle large datasets efficiently (Fahad et al., 2014).

The non-hierarchical methods mentioned above are most suitable to either handle numerical or categorical attributes. However, objects encountered in real world databases are often *mixed-types of data*. Huang (1998) integrated the k-means and k-modes algorithms and introduced the *k-prototypes* algorithm that can be used to cluster mixed-type objects and is capable to handle large datasets and high dimensionality. The algorithm includes the squared Euclidean distance measure for numeric attributes and the simple matching dissimilarity measure for categorical attributes (Huang, 1998). A certain weight is used to avoid favouring a type of attribute whereby the researcher's knowledge about the data is an important factor.

For *hierarchical clustering* various algorithms are available in literature. Guha, Rastogi, and Shim (1998) introduced and applied the hierarchical algorithm *CURE* for clustering large datasets. The algorithm considers the scattered points as representatives to capture the shape and extent of the cluster (Guha et al., 1998). The closest pair of representative points are merged at each step to generate the clusters. According to Guha (1998) and Fahad et al. (2014) it can not only handle large datasets but also high dimensionality and it is more robust against noise because shrinking the scattered points toward the mean reduces sensitivity to outliers. However, it is applicable on numerical data only (Fahad et al. 2014). Karypis, Han, and Kumar (1999) introduced and applied the hierarchical algorithm *Chameleon* which is based on dynamic modelling. A key feature is that it considers the interconnectivity and closeness in identifying the most similar pair of clusters (Karypis, 1999). Hence, two clusters are merged when the interconnectivity and proximity (closeness) between clusters is high compared to the within cluster interconnectivity and closeness of objects. Karypis et al. (1999) states that as long as a similarity matrix can be provided, the dynamic modelling of clusters in the Chameleon algorithm is applicable to all types of data, handling large datasets, and high dimensionality. Guha et al. (2000) introduced the *ROCK* algorithm which is applicable to both numerical and categorical variables (Guha et al., 2000; Fahad et al., 2014). As argued in Guha et al. (2000) the ROCK algorithm uses a *links-based measure* and not a distance-based measure as a basis to merge neighbouring data points to create clusters. While the ROCK algorithm is capable of handling large datasets, it is less efficient at handling high dimensionality or noise (Guha et al., 2000; Fahad et al. 2014). Tamasauskas, Sakalauskas, & Kriksciuniene (2012) evaluated the performance of ten different hierarchical clustering methods by experimenting with ten different similarity measures in terms of accuracy. The study considered hierarchical methods including single linkage, complete linkage, average linkage, centroid's method, density linkage, flexible-beta, McQuitty's, median, two-stage density linkage, and Ward's method. Performance evaluation revealed that the best algorithms are *complete linkage, Ward's method, and flexible-beta* (Tamasauskas et al., 2012). However, the latter hierarchical clustering methods are computationally expensive and slow when handling large datasets and high dimensionality compared to the Chameleon, ROCK, and CURE algorithms (Fahad et al., 2014).

In addition to hierarchical and non-hierarchical methods the *model-based* method is often used in academic literature for clustering. Dolnicar (2002) and Fahad et al. (2014) mentioned Neural Networks became a more prevalent application in literature for clustering solutions. According to Santana et al. (2017) the Self-Organising Maps (SOMs) algorithm introduced by Kohonen (1998) is the most used type of neural network. SOMs can provide models for

clustering, classification, and forecasting (Sathya, & Abraham, 2013). The goal of SOMs is to convert an input signal (high dimensional) into a simpler discrete map (Larose, 2014). Additionally, it used for data visualization or dimensionality reduction purposes (Kohonen, 2013). SOMs structures output nodes into clusters of nodes where nodes in closer proximity are more similar than to other nodes that are further apart (Larose, 2014; Kohonen 2013). SOMs are less sensitive to initialization and it is not required to specify the number of clusters a priori (Murray, Agard, & Barajas, 2017).

3.1.5 Two-Stage Clustering and Data Size

Determining the number of clusters a priori most strongly influences clustering solutions. The problem of selecting the number of clusters is one of the oldest unsolved problems in clustering analysis (as cited in Dolnicar, 2002). One of the first approaches were suggested by Milligan (1981) and Milligan & Cooper (1985) which are based on an internal index comparison. However, a two-stage clustering methodology was proposed by Punj and Stewart (1983) wherein they recommended to identify clusters by first using Ward's method or average linkage (i.e., hierarchical clustering) followed by non-hierarchical clustering for cluster refinement. They concluded a two-stage approach yields better results than solely using a hierarchical or non-hierarchical approach. Mazanec and Strasser (2000) adopted a two-stage approach of hierarchical and non-hierarchical clustering and drew similar conclusions of obtaining superior results. Kuo, Ho, and Hu (2002) modified the two-stage approach and proposed to use self-organising maps (i.e., model-based) to determine the number of clusters followed by the k-means algorithm. They concluded their modified two-stage method provided good solutions for determining the initial segments and observed a reduced number of misclassifications compared to conventional methods. Hence, determining the number of clusters by hierarchical clustering before applying a non-hierarchical procedure is an appropriate method for obtaining robust clustering results.

Hierarchical clustering methods are computationally expensive and slow when handling large datasets or high dimensionality (Fahad et al., 2014). Therefore, literature is reviewed in order to provide some indications on what data size could be referred to as too large or small. Generally, non-hierarchical methods have superior performance on large data sets whereas the performance of hierarchical methods decreased as the number of observations increased (Zhao & Karypis, 2002; Abbas, 2008). Dolnicar (2002) studied the standards of clustering analysis in academic literature for data-driven market segmentation and found that the smallest data size contained only 10 objects, the largest 20.000 objects, and the average size was 700. In case of hierarchical clustering methods the data sizes contained 530 observations on average and for non-hierarchical methods 927. The number of variables in the datasets ranged between 10 and 66 variables, with a mean number of 17 variables (Dolnicar, 2002; Dolnicar, 2003). Therefore, one could potentially regard 10 variables as low dimensionality and more than 10 variables as high dimensionality. Other studies have applied hierarchical clustering methods on varying data sizes. For instance, Abbas (2008) evaluated the performance of hierarchical and non-hierarchical clustering methods on data sizes of 4000 and 36000 with varying dimensionality and numbers of clusters. Results indicated that hierarchical clustering performed best on a smaller dataset with low dimensionality. Therefore, a data size of less than 4000 observations could potentially be considered as being small enough for hierarchical clustering and its computation time. Datasets with more than 4000 observations could be considered as large and potentially less suitable for hierarchical clustering methods except for the Chameleon, ROCK, and CURE algorithms. Due to a lack of rules regarding the data size, the only recommendation that could be given is to question if the dimensionality is not too high for the number of cases to be grouped (Dolnicar, 2002; Dolnicar, 2003). One approach to determine the minimum data size is to include no less than 2^k cases (k = number of variables), and preferably $5*2^k$ (Dolnicar, 2002).

In brief, hierarchical clustering is applicable when more than one clustering solution is of interest or the data size is moderate. The number of clusters can be determined by hierarchical clustering and a non-hierarchical procedure then clusters all observations using the determined number of clusters or initial seed points to provide more accurate cluster memberships.

3.2 Framework for User Profiling based on Unsupervised Machine Learning

A framework is proposed to visualize User Profiling strategies based on Unsupervised Machine Learning (UML) and the requirements and characteristics of the dataset. The framework is based on literature discussed in chapter 3.1. Selecting a particular algorithm for UML problems is highly dependent on the *data type, data size, and data dimensionality*. These data properties have a significant effect on the quality and efficiency of the clustering procedure and solution (Fahad et al., 2014, Pandove et al., 2018, Dolnicar., 2002). For instance, when analysing a large numerical dataset one might apply k-means and for large categorical data k-modes.

However, Dolnicar (2002) studied the standards of clustering analysis in academic literature for data-driven market segmentation and found that the smallest data size contained only 10 objects, the largest 20.000 objects, and the average size was 700. The number of variables in the datasets ranged between 66 and 10 variables, with a mean number of 17 variables (Dolnicar, 2002; Dolnicar, 2003). Therefore, one could potentially regard 10 variables as *low dimensionality* and more than 10 variables as *high dimensionality*. Additionally, Abbas (2008) evaluated the performance of hierarchical and non-hierarchical clustering methods on data sizes of 4000 and 36000 with varying dimensionality and numbers of clusters. Results indicated that hierarchical clustering performed best on a smaller dataset with low dimensionality. Therefore, a data size of less than 4000 could potentially be considered small enough for hierarchical clustering and its computation time and interpretability. Datasets with more than 4000 observations can be considered as large and potentially less suitable for hierarchical clustering methods except for the Chameleon, ROCK, and CURE algorithms (Section 3.1.5). The assumptions mentioned above provide a rough estimation about what could be considered as high or low dimensionality and large or small data sizes. However, they remain to be assumptions and a lack of rules exist regarding these categorizations in academic literature. According to Dolnicar (2002) the only recommendation that could be given is to question if the dimensionality is not too high for the number of cases to be grouped (i.e., 2^k cases and preferably $5*2^k$). Table 1 provides an overview of the clustering algorithms with respect to the data characteristics as described in section 3.1.4.

The *Framework* in Table 2 outlines various strategies for User Profiling based on unsupervised machine learning and the data properties including the data type, data size, and dimensionality. The framework includes strategies for categorical, numerical, and mixed types of data. The first stage consists of an hierarchical or model-based clustering procedure to determine the number of clusters and identify initial seeds. Secondly, a non-hierarchical clustering procedure is applied to provide more accurate cluster memberships.

Table 1

Overview of clustering algorithms and data characteristics as reviewed in section 3.1.4

Category	Algorithm	Data Type	Data Size	Handling High Dimensionality	Handling Noise
Model-Based Algorithms	SOMs (Kohonen, 1998)	Multivariate Data	Small/Moderate	Yes	No
Hierarchical Algorithms	Chameleon (Karypis et al., 1998)	Categorical/Numerical	Large	Yes	No
	ROCK (Guha et al., 2000)	Categorical/Numerical	Large	No	No

	CURE (Guha et al., 1998)	Numerical	Large	Yes	Yes
	Complete Linkage/Ward's (Tamasauskas et al., 2012; Pandove et al., 2018;)	Dependent on Distance Measure	Small/Moderate	No	No
Non-Hierarchical Algorithms	K-modes (Huang, 1998)	Categorical	Large	Yes	No
	K-medoids (Park et al., 2009)	Categorical	Small	Yes	Yes
	K-means (MacQueen, 1967)	Numerical	Large	No	No
	K-prototypes (Huang, 1998)	Categorical/Numerical	Large	Yes	No

Note. Adapted from Fahad et al. (2014)

Table 2

Framework outlining Unsupervised Machine Learning algorithms for User Profiling based on Two-Stage clustering and the characteristics of the dataset

Data Type	Data Size	Dimensionality	Stage - 1	Stage - 2
Categorical	Large	High	Chameleon	K-modes
		Low	ROCK	K-modes
	Small/Moderate	High	Chameleon	K-modes/K-medoids
		Low	Complete Linkage/Ward's	K-modes/K-medoids
Numerical	Large	High	CURE	K-means
		Low	CURE	K-means
	Small/Moderate	High	SOMs	K-means
		Low	SOMs	K-means
Categorical/Numerical (Mixed)	Large/Small	High	Chameleon	K-prototypes
		Low	ROCK	K-prototypes

Note. A data size of <=4000 is considered to be moderate/small. High Dimensionality is approximately >10 variables and Low Dimensionality is <=10 variables. A lack of rules exists regarding these data properties in literature (see section 3.1.5).

4. DISCUSSION

The purpose of this paper was to develop a methodology and a framework of Unsupervised Machine Learning (UML) algorithms with respect to important data properties for the purpose of User Profiling and data-driven customer segmentation. A key competitive advantage for today's organizations is the availability of large amounts of data for the purpose of segmenting a customer base, offering tailored services, and extracting meaningful information provided by various data sources. However, organizations often have difficulties to extract knowledge from data and selecting appropriate Machine Learning algorithms in order to develop accurate User Profiles and segments. Moreover, marketing departments often lack a fundamental understanding on data-driven segmentation methodologies. Key issues were determining the number of clusters and which algorithm should be chosen. In addition, numerous approaches were available for numerical data but approaches for categorical or mixed data were not as prevalent and straightforward. Prior research focused on the development, effectiveness (i.e., accuracy), and efficiency of various UML algorithms (e.g., Tamasauskas et al., 2012; Pandove et al., 2018, Huang, 1998; Park et al., 2009). However, none of the studies provided an outline of UML algorithms and the various requirements and characteristics regarding the dataset. The research question was as follows: *What is an appropriate framework for outlining Unsupervised Machine Learning Algorithms for User Profiling?* Literature was reviewed

regarding the core concepts of UML and various algorithms, two-stage clustering, and the characteristics and requirements regarding the data properties.

A *framework* is proposed outlining various Unsupervised Machine Learning algorithms for User Profiling with respect to various data properties. It provides a two-stage clustering methodology for categorical, numerical, and mixed types of data with respect to the data size and data dimensionality. The first stage consists of an hierarchical or model-based clustering procedure to determine the number of clusters. In the second stage, a non-hierarchical clustering procedure is applied for cluster refinement.

The framework contributes to body of knowledge regarding approaches and methodologies for UML and data-driven segmentation in a marketing context. Until now, none provided an outline consisting of a two-stage clustering approach for UML algorithms, different types of data, and various characteristics of the dataset. The two-stage clustering approach alleviates the drawbacks of solely using hierarchical or non-hierarchical clustering procedures which can result in more robust clustering solutions (Kuo et al., 2002; Mazanec & Strasser, 2000; Punj & Steward, 1983).

Practical implications are that the framework can support researchers and practitioners to determine which UML algorithms are appropriate for developing robust user profiles and data-driven customer segments for marketing purposes.

5. REFERENCES

Abbas, A. (2008). Comparisons between data clustering algorithms. *International Arabic Journal of Information technology, 5*(3), 320-325.

Alamuri, M., Surampudi, B. R., & Negi, A. (2014). A survey of distance/similarity measures for categorical data. *Neural Networks International Joint Conference*, 1907-1914. IEEE

Boriah, S., Chandola, V., & Kumar, V. (2008). Similarity Measures for Categorical Data: A Comparative Evaluation. *Proceedings of the 2008 SIAM Conference on Data Mining*, 243-254.

Boratto, L., Carta, S., Fenu, G., & Saia, R. (2016). Using neural word embedding's to model user behavior and detect user segments. *Knowledge-based systems, 108*, 5-14. https://doi.org/10.1016/j.knosys.2016.05.002

Bose, I., & Mahapatra, R. K. (2001). Business data mining – a machine learning perspective. *Information & Management, 39*(3), 211-225.

Chester, J. (2012). Cookie Wars: How New Data Profiling and Targeting Techniques Threaten Citizens and Consumers in the "Big Data" Era. *European Data Protection: In Good Health?* 53-77. doi:10.1007/978-94-007-2903-2_4

Dolnicar, S. (2008). Market segmentation in tourism. *Tourism management, analysis, behaviour and strategy*, 129-150. Retrieved from https://pdfs.semanticscholar.org/d7d0 /1f681371015892e18f c7f68ea9a1dbd878bd.pdf

Dolnicar, S. (2003). Using cluster analysis for market segmentation - typical misconceptions, established methodological weaknesses and some recommendations for improvement. *Australasian Journal of Market Research, 2003, 11*(2), 5-12.

Dolnicar, S., & Lazarevski, K. (2009). Methodological reasons for the theory/practice divide in market segmentation. *Journal of marketing management, 25*(3-4), 357-373.

Eirinaki, M., & Vazirgiannis, M. (2003) . Web mining for web personalization. *ACM Transactions on Internet Technology (TOIT), 3*(1), 1-17.

Fahad, A., Alshatri, N., Tari, Z., Alamri, A., Khalil, I., Zomaya, A. Y., & Bouras, A. (2014). A survey of clustering algorithms for big data: Taxonomy and empirical analysis. *IEEE transactions on emerging topics in computing, 2*(3), 267-279.

Guha, S., Rastogi, R., & Shim, K. (2000). ROCK: A robust clustering algorithm for categorical attributes. *Information Systems, 25*(5), 354-366. https://doi.org/10.1016/S0306-4379(00)00022-3

Guha, S., Rastogi, R., & Shim, K. (1998). CURE: an efficient clustering algorithm for large databases. *AMC Sigmod Record, 27*(2), 73-84. Retrieved from https://dl-acm-org.ezproxy2.utwente.nl/citation.cfm?doid=276305.276312

Han, J., Kamber, M., & Pei, J. (2012). Cluster Analysis Concepts and Methods. *Data Mining, 443-495*. Retrieved from: https://doi.org/10.1016/B978-0-12-381479-1.00010-1

Hiziroglu, A. (2013). Soft computing applications in customer segmentation: State-of-art review and

critique. *Expert Systems with Applications, 40*(16), 6491-6507.
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. https://doi.org/10.1023/A:1009769707641
Jordan, M. I., & Mitchell, T.M. (2015). Machine learning: trends, perspectives, and prospects. *Science, 349*(6245), 255-260.
Karypis, G., Han, E. H., & Kumar, V. (1999). Chameleon: Hierarchical clustering using dynamic modeling. *Computer, 32*(8), 68-75.
Kohonen, T. (2013). Essentials of the self-organizing map. *Neural networks, 37*(9), 52-65. doi: https://doi.org/10.1016/j.neunet.2012.09.018
Kohonen, T. (1998) The Self-Organizing Map. *Neurocomputing, 21*, 1-6. http://dx.doi.org/10.1016/S0925-2312(98)00030-7
Kuo, R. J., Ho, L. M., & Hu, C.M. (2002). Cluster analysis in industrial market segmentation through artificial neural networks. *Computers & Industrial Engineering, 42*(2-4), 391-399.
Larose, D. T. (2014). Discovering knowledge in data: introduction to data mining. John Wiley & Sons.
Pandove, D., Goel, S., & Rani, R. (2018). Systematic review of clustering high-dimensional and large datasets. *AMC transactions on knowledge discovery from data (TKDD), 12*(2),1-68. doi: https://doi.org/10.1145/3132088
Park, H.S., & Jun, C. H. (2009). A simple and fast algorithm for K-medoids clustering. *Expert systems with applications, 36*(2), 3336-3341. doi:10.1016/j.eswa.2008.01.039
Prasad, Y.L. (2016). Big data analytics made easy. USA: Notion Press.
Punj, G., & Stewart, D. (1983). Cluster analysis in marketing research: Review and suggestions for application. *Journal of marketing research*, 134-148.
MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of the fifth Berkeley symposium on mathematical statistics*, *1*(14), 281-297.
Malhotra, N. K. (2004) *Marketing research: an applied orientation, 4th edition,* Prentice-Hall International, London.
Mazanec, J. A., & Strasser, H. (2000). A nonparametric approach to perceptions-based market segmentation: Foundations, (1). Springer
Milligan, G. W. (1981). A review of Monte Carlo tests of cluster analysis. *Multivariate behavioural research, 16*(3), 379-407.
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.
Murray, P. W., Agard, B., & Barajas, M. A. (2017). Market segmentation through data mining: A method to extract behaviors from a noisy data set. *Computers & Industrial Engineering, 109*, 233-252. doi: http://dx.doi.org/10.1016/j.cie.2017.04.017
Santana, A., Morais, A., & Quiles, M. G. (2017). An alternative approach for binary and categorical self-organizing maps. In *Neural Networks (IJCNN), 2017 International Joint Conference on* (pp. 2604-2610). IEEE.
Sathya, R., & Abraham, A. (2013). Comparison of supervised and unsupervised learning algorithms for pattern classification. *International Journal of Advanced Research in Artificial Intelligence, 2*(2), 34-38. Doi: 10.1109/TITB.2012.2223823
Tamasauskas, D., Sakalauskas, V., & Kriksciuniene, D. (2012). Evaluation framework of hierarchical clustering methods for binary data. *Hybrid Intelligent Systems* (HIS), 2012 12th International Conference, 421-426. Doi: 10.1109/ICHPCA.2014.7045336
Lorenco, F., Lobo, V., & Bacao, F. (2004). Binary-based similarity measures for categorical data and their application in Self-Organizing maps, 1-18.
Turing, A. M., (1950). Computing machinery and intelligence. *Mind 59*(236), 433–460.
Webster, J., & Watson, R. T. (2002). Analyzing the past to prepare for the future: Writing a literature review. MIS quarterly.
Walters, M., & Bekker, J. (2017). Customer super-profiling demonstrator to enable efficient targeting in marketing campaigns. *South African Journal of Industrial engineering, 28*(3), 113-127. Retrieved from sajie.journals.ac.za/pub/article/download/1846/807
Wolfswinkel, J.F., Furtmueller, E., & Wilderom, C.P. (2013). Using grounded theory as a method for rigorously reviewing literature. *European journal of information systems, 22*(1), 45-55.
Zhao, Y., & Karypis, G. (2002). Evaluation of hierarchical clustering algorithms for document datasets. *International conference on information and knowledge management*, 515-542.