DyClee-N&C: a clustering algorithm for heterogeneous data based situation assessment

Audine Subias, Louise Travé-Massuyès, Tom Obry

DyClee-N&C: a clustering algorithm for heterogeneous data based situation assessment ⋆

Audine Subias[*,**], **Louise Travé-Massuyès**[*,**], **Tom Obry** [*]

[*] *LAAS-CNRS, Université de Toulouse, CNRS, INSA, UPS, Toulouse, France, (subias, louise@laas.fr)*
[**] *ANITI, Université Fédérale Toulouse Midi Pyrénées, France*

Abstract: In data-based situation assessment applications, the proliferation of data acquired and recorded on current technological systems is a key issue in that data remain unlabeled because labeling would require too much time and implies prohibitive costs. The data should therefore speak for itself. The different situations, e.g., normal or faulty, must hence be learned only from the data. Clustering methods, also named unsupervised classification methods, can be used for that purpose. These methods are designed to cluster the samples according to some similarity criterion. The different clusters can be associated to different situations whose discrimination may be relevant to obtain a proper diagnosis.
Numerous algorithms have been developed in recent years for clustering numeric data but these methods are not applicable to categorical data. This is the case of the algorithm *DyClee*, named *DyClee-N* in the paper. However, in many application domains, qualitative features are key to properly describe the different situations. *DyClee-N* was recast to produce a version, named *DyClee-C* that accepts categorical features, but only categorical features. This paper presents *DyClee-N&C* that subsumes both the numeric and categorical feature based algorithms *DyClee-N* and *DyClee-C* respectively. *DyClee-N&C* is applied to a data set of the literature for the evaluation of risk in the automobile domain and compared to state of the art clustering methods.

Keywords: Dynamic clustering, Situation assessment, artificial intelligence.

1. INTRODUCTION

In the digital age, the amount of data that are recorded by organizations and companies is enormous. If these data are to have added value, it must be possible to extract relevant information automatically. This is why data mining methods appear to be crucial. Among them, clustering methods have an essential role to play. Indeed, data often remain unlabelled because labelling would require too much time and imply prohibitive costs. In diagnosis applications for instance, the different situations, e.g. normal or faulty, must hence be learned from the data. Clustering methods, also qualified as unsupervised classification methods, can then be used to create groups of samples according to some similarity criterion. The different groups can supposedly be associated to different situations.

In the field of *clustering*, several methods have been developed to group together data composed solely of numerical or categorical features. However, few algorithms provide a way to partition similar samples described by mixed features. Some methods have been proposed such as the *K-Prototype* (Huang, 1998), *ClustMD* (McParland and Gormley, 2016), *CluMix* (Hummel et al., 2017), *CEBMDC* (He et al., 2005) or even *Fuzzy K-Prototype* (Ji et al., 2012). All these algorithms use their own notion of similarity to create clusters. Indeed, classical numeric distance metrics such as the Euclidean distance or the Manhattan distance for high dimensional data sets are not applicable to categorical data. Conversely, metrics for assessing the similarity of samples described by categorical features are not applicable to numerical data. However, a mixture of numeric and qualitative features is often required to properly describe the different objects/situations in many application domains.

In this paper, we leverage two versions of the same algorithm, the original DyClee (**Dy**namic **Cl**ustering algorithm for tracking **E**volving **E**nvironments)(Barbosa Roa et al., 2019) that only accepts numeric features and an extension to categorical features named *DyClee-C*(Obry et al., 2019). DyClee features several properties like handling non convex, multi-density clustering with outlier rejection, and it achieves full dynamicity. All these properties are not generally found together in the same algorithm.

In this paper, Dyclee is renamed *DyClee-N* for better understanding. *DyClee-N* and *DyClee-C* are the building blocks of the mixed numeric/qualitative version presented in this paper named *DyClee-N&C*. *DyClee-N&C* hence represents a versatile and usable dynamic clustering algorithm for data described by all types of features. This gives it a definite advantage for some situation assessment and

⋆ This work is part of the CIFRE PhD project in collaboration with the company ACTIA. It is also related to ANITI within the French "Investing for the Future – PIA3" program under the Grant agreement n°ANR-19-PI3A-0004.

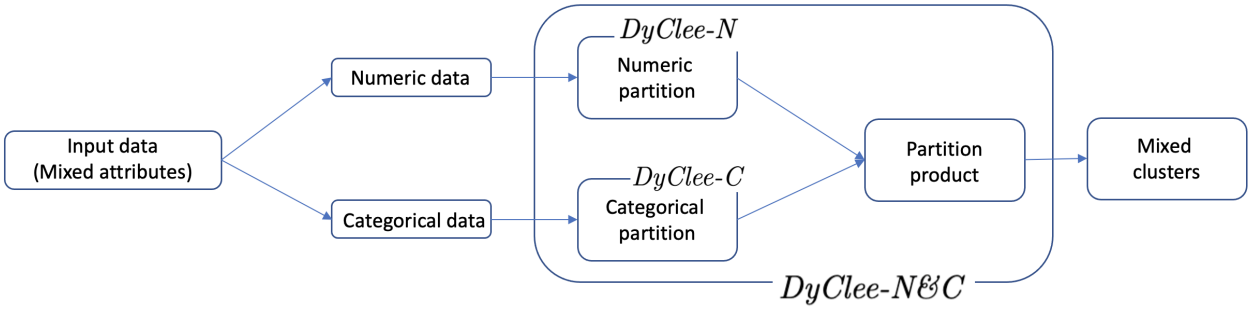

Fig. 1. Principles of *DyClee-N&C*

diagnosis applications in which the goal is to discriminate normal and the different faulty states.

The paper is organized as follows. Section 2 presents the basic principles and the different steps of the *DyClee-N&C* algorithm. *DyClee-N* and the qualitative extension *DyClee-C* are introduced in sections 2.1 and 2.2. The integration of *DyClee-N* and *DyClee-C* is then presented in section 3. The tests on a public data set are presented in section 4. Finally, section 5 provides conclusions and perspectives of this work.

2. PRINCIPLE OF *DYCLEE-N&C*

DyClee-N&C is proposed as a dynamic clustering algorithm that integrates both the numeric and categorical versions *DyClee-N* and *DyClee-C*. The samples $\mathcal{X} = \{\mathcal{X}_1, \mathcal{X}_2, ..., \mathcal{X}_i, ..., \mathcal{X}_n\}$ as input to the algorithm are described by numeric and categorical features in the set $\mathcal{A} = \{A_1, A_2, ..., A_d, \mathbb{A}_1, \mathbb{A}_2, ..., \mathbb{A}_d\}$, where d is the number of numeric features and d the number of categorical features. The data set obtained by projecting $\mathcal{X}$ onto numeric features and categorical features are denoted X and $\mathbb{X}$, respectively. The principle of *DyClee-N&C* follows two steps:

(1) cluster X, i.e. the samples with respect to the numeric features only, and $\mathbb{X}$, i.e. the samples with respect to the categorical features only and obtain two partitions P and $\mathbb{P}$,
(2) Merge the two partitions to obtain the final clustering.

The first step relies on *DyClee-N* and *DyClee-C* that are presented in sections 2.1 and 2.2, respectively. The merging step, which achieves *DyClee-N&C* is presented in section 3. Figure 1 gives the overview of mixed *DyClee-N&C*.

Note that the output of a clustering is a partition, as defined formally below, of the set of samples provided as input.

Definition 2.1. A partition is defined as a set of subsets $\mathcal{P} = \{\mathcal{P}_1, \mathcal{P}_2, ..., \mathcal{P}_k\}$ of a given set S with:

- $\emptyset \notin S$
- $\mathcal{P}_1 \cup \mathcal{P}_2 \cup ... \cup \mathcal{P}_k = S$
- $\mathcal{P}_i \cap \mathcal{P}_j = \emptyset$ with $i,j = 1,...,$ k and $i \neq j$.

2.1 *NUMERIC CLUSTERING WITH* DyClee-N

DyClee-N integrates a distance and a density based algorithm.

The first step qualified as *distance-based step* operates at the rate of the data stream and creates micro-clusters (μ-clusters), putting together data samples that are close in the sense of the L1-norm. μ-clusters are stored in the form of summarized representations including statistical and temporal information gathered in a characteristic vector. Centers of μ-clusters can be calculated by averaging, for each feature, all the samples present in the μ-cluster. The reachability of μ-cluster from a data sample is evaluated based on the Chebyshev distance and the sample is assigned to the closest μ-cluster according to the Manhattan distance.

The second step, qualified as *density-based step*, gathers the μ-clusters to create the final clusters. A cluster is defined as a set of "connected" μ-clusters, where every inside μ-cluster presents high density and every boundary μ-cluster exhibits either medium or high density. Figure 2 illustrates how the *DyClee-N* algorithm works.

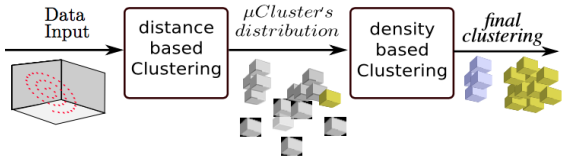

Fig. 2. Overview of *DyClee-N*

In *DyClee-N*, clusters are created at the beginning of the density-based step from groups of μ-clusters found by the KD-Tree algorithm (Maneewongvatana and Mount, 1999). A KD-Tree is a space-partitioning data structure for organizing points in a k-dimensional space. In *DyClee-N*, the K-tree algorithm is used to efficiently find the nearest neighbors of μ-clusters and form μ-clusters groups. Then, for each group, the densest μ-cluster subregions are identified. A cluster is created if a μ-cluster is dense and if its neighbors are dense or semi-dense inside a group. If a μ-cluster is outlier, all the samples in this μ-cluster are considered noise.

DyClee-N offers two approaches to qualify the density of a μ-cluster and then to find the clusters: the global approach and the local approach.

In the global approach, two global density thresholds are defined as the median and average densities of all μ-clusters. A μ-cluster μC_z with density D_z is said to be:

- *dense* if D_z is greater than or equal to both thresholds,
- *semi-dense* if D_z is greater or equal to one of the two thresholds and lower than the other,
- *outlier* if D_z is strictly less than the two thresholds.

In the local approach, the two global density thresholds are replaced by two local density thresholds that are the median and average densities of the μ-clusters of the group to which μC_z belongs. The local approach is very useful for multidensity data sets.

DyClee-N also implements a forgetting process in order to follow the data evolution at best. Data in μ-clusters are subject to a decay function at the beginning of the density-based step.

2.2 CATEGORICAL CLUSTERING WITH DyClee-C

DyClee-C follows the same principles as *DyClee-N* with the adaptations required to deal with qualitative data instead of numeric data. Like *DyClee-N* it is based on both a distance step and a density step.

The definition of μ-cluster is adapted by considering the frequency of the qualitative features modalities to determine its *center*. Let $\mathbb{A} = \{\mathbb{A}_1, \mathbb{A}_2, ..., \mathbb{A}_\mathrm{d}\}$, be the set of qualitative features characterizing the samples, with d the number of features. The set of modalities of the kth feature is denoted $Mod(\mathbb{A}_k) = \{\mathrm{m}_1^k, \mathrm{m}_2^k, ..., \mathrm{m}_{p_k}^k\}$, with p_k the number of modalities. Consider a qualitative μ-cluster $\mu\mathbb{C}_i$, then the frequency of the most frequent modality in $\mu\mathbb{C}_i$ relatively to the feature k is given by:

$$\mathbb{F}_i^k = \arg\max_{\mathrm{m}_j^k}(fr_i(\mathrm{m}_j^k)), j \in \{1, ..., p_k\} \quad (1)$$

where $fr_i(\mathrm{m}_j^k)$ is the frequency of the j^{th} modality of the k^{th} feature in the samples of $\mu\mathbb{C}_i$. The vector $\mathbb{F}_i = \{\mathbb{F}_i^1, \ldots, \mathbb{F}_i^\mathrm{d}\}$ gathers the most frequent modalities in $\mu\mathbb{C}_i$ in all qualitative dimensions and it identifies the qualitative center of $\mu\mathbb{C}_i$. Dealing with qualitative data also enforces two main changes:

- In the distance-based step: the distance used to assign samples to μ-clusters and to assess reachability of μ-clusters from data samples is now taken as the Hamming distance.
- In the density-based step, the Locality Sensitive Hashing (LSH) algorithm ((Indyk and Motwani, 1998) (Leskovec et al., 2014), (Gionis et al., 1999)) replaces the KD-Tree algorithm that does not accept qualitative data.
  The global and local density approaches defined in *DyClee-N* to evaluate the density of a μ-cluster are similar in *DyClee-C* but the density is replaced by $\mathbb{D}_i$, the number of samples in $\mu\mathbb{C}_i$.

Like *DyClee-N* *DyClee-C* also implements a forgetting process to follow the data dynamics. The μ-clusters characteristic vector are revised by considering the age of the included samples at the beginning of the density-based step.

3. *DYCLEE-N&C*

This section presents the algorithm *DyClee-N&C* that achieves an integration of *DyClee-N* and *DyClee-C* and therefore accepts numeric and qualitative features as input.

3.1 Merging numeric and categorical partitions

Clustering the samples with respect to the numeric features with *DyClee-N* and with respect to the qualitative features with *DyClee-C*, provides two partitions P and $\mathbb{P}$, called the numeric and the categorical partition respectively. The second step of *DyClee-N&C* is to merge these two partitions. To do so, we propose to make the product of the numeric partition P and the qualitative partition $\mathbb{P}$, which would result in a partition, i.e., a clustering, accounting for both numeric and qualitative features.

A partition product applies to two partitions of the same set. However, the μ-clusters formed from numerical and from categorical features are not the same since they do not necessarily group together the same samples. We must therefore place ourselves at the level of the samples. Note also that, in the default mode (other parameterizations are presented in Section 3.4), *DyClee-N* and *DyClee-C* reject outlier μ-clusters. These latter do not necessarily gather the same samples on both sides. Our proposal is therefore to discard the samples assigned to outlier μ-clusters on either side.

Figure 3 illustrates the outlier removal stage with six μ-clusters, three numeric μ-clusters μC_1, μC_2, and μC_3, and three qualitative μ-clusters $\mu\mathbb{C}_1$, $\mu\mathbb{C}_2$, and $\mu\mathbb{C}_3$. Samples are numbered from 1 to 10 and they have been assigned to μ-clusters on each side. The μ-clusters considered as *dense* are represented in red color, the *semi-dense* μ-clusters are represented in orange color and the grey color is given to *outlier* μ-clusters. Sample 4 is the only sample assigned to an outlier μ-cluster which is in this case the qualitative μ-cluster $\mu\mathbb{C}_1$. Following our proposal, sample 4 is hence removed from each side, leaving μC_1 with samples 1, 2, and 3 on the numeric side and discarding $\mu\mathbb{C}_1$ as a whole on the qualitative side. The set of samples is then the same on both sides.

We adopt the following assumption.

Assumption: Removing the samples that belong to outlier μ-clusters in one of the partitions, numerical or categorical, does not change the density status to outlier for the non-outlier μ-clusters to which the samples belong in the other partition.

The above assumption means that new outlier μ-clusters are not created in the removal phase, which can hence be done in a single pass. It can be justified from the specific density thresholds that have been defined (cf. Section 2).

The set of samples being the same on both the numeric and the categorical side, we then proceed to the final density based clustering stage on both sides. We obtain the numeric and categorical partitions that, by the assumption above, partition the same set. The product of partitions as defined below can then be made.

Definition 3.1. The product of two partitions $\mathcal{P}^A = \{\mathcal{P}_1^A, \ldots, \mathcal{P}_{n_A}^A\}$ and $\mathcal{P}^B = \{\mathcal{P}_1^B, \ldots, \mathcal{P}_{n_B}^B\}$ of the same set S is a partition $\mathcal{P} = \{\mathcal{P}_1, \ldots, \mathcal{P}_n\}$ defined by:

$$\forall x \forall y \in S \quad x \in \mathcal{P}_k, y \in \mathcal{P}_k, k \in \{1, \ldots, n\} \Leftrightarrow$$
$$\exists i \in \{1, \ldots, n_A\}, \exists j \in \{1, \ldots, n_B\}, x \in \mathcal{P}_i^A,$$
$$y \in \mathcal{P}_i^A, x \in \mathcal{P}_j^B, y \in \mathcal{P}_j^B.$$

Definition 3.1 states that the product of two set partitions $\mathcal{P}^A$ and $\mathcal{P}^B$ is defined as the set partition whose parts are the nonempty intersections between each part of $\mathcal{P}^A$ and each part of $\mathcal{P}^B$.

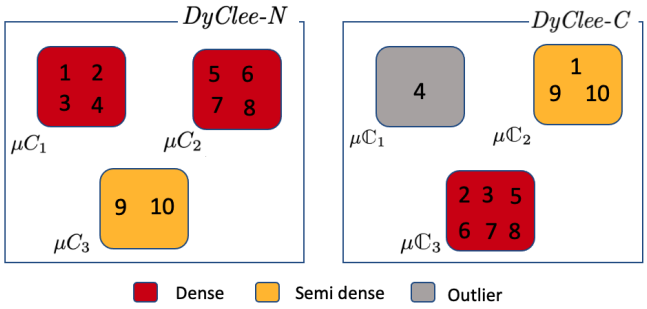

Fig. 3. The *outlier* problem

3.2 DyClee-N&C *algorithm*

The different steps of *DyClee-N&C* are listed below:

(1) Run in parallel the distance-based step of *DyClee-N* and *DyClee-C* to forms numeric and categorical μ-clusters considering the samples X and $\mathbb{X}$ respectively.
(2) Label numeric and categorical μ-clusters by dense, semi-dense, or outlier according to the global or local density thresholds.
(3) Remove the samples from outlier μ-clusters on each side.
(4) Run in parallel the density-based step of *DyClee-N* and *DyClee-C* to obtain the numeric and the categorical partitions P and $\mathbb{P}$.
(5) Make the product of the numeric and categorical partitions P and $\mathbb{P}$ to obtain a partition that stands for the final clustering counting for numeric and categorical features.

The *DyClee-N&C* algorithm also adapts to the situation where the input data set has only numeric or categorical features. In such cases, the appropriate version of *DyClee* (numeric or categorical) is run alone.

3.3 Illustrative example

The final step (5) of *DyClee-N&C* is illustrated with the example of Figure 4.

Consider that steps (1) to (4) have resulted in the numeric partition of three sets, or clusters, of numeric μ-*clusters* $P = \{Cl_1{=}[\mu C_1, \mu C_2], Cl_2{=}[\mu C_3, \mu C_4], Cl_3{=}[\mu C_5]\}$ and the categorical partition composed of two sets of categorical μ-*clusters* $\mathbb{P} = \{\mathbb{C}l_1{=}[\mu\mathbb{C}_1, \mu\mathbb{C}_2], \mathbb{C}l_2{=}[\mu\mathbb{C}_3]\}$.

The numeric μ-clusters contain the following samples: $\mu C_1 = \{1, 2, 3\}$, $\mu C_2 = \{4, 5\}$, $\mu C_3 = \{6, 7\}$, $\mu C_4 = \{8, 9\}$ and $\mu C_5 = \{10, 11, 12\}$. On the other hand, categorical μ-clusters contain the following samples: $\mu\mathbb{C}_1 = \{1, 2, 3\}$, $\mu\mathbb{C}_2 = \{4, 5, 6\}$ and $\mu\mathbb{C}_3 = \{7, 8, 9, 10, 11, 12\}$.

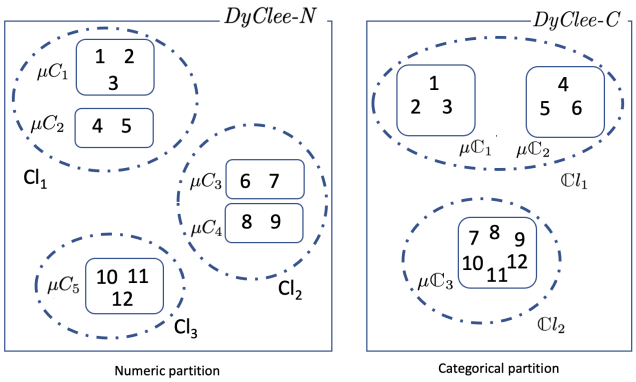

Fig. 4. Example of partitions of numeric and categorical μ-*clusters*

Placing ourselves at the level of samples, the partitions P^A and P^B of Definition 3.1 correspond respectively to the partitions $P_s = \{[1,2,3,4,5], [6,7,8,9], [10,11,12]\}$ and $\mathbb{P}_s = \{[1,2,3,4,5,6], [7,8,9,10,11,12]\}$ derived from P and $\mathbb{P}$. This is illustrated by Figure 5.

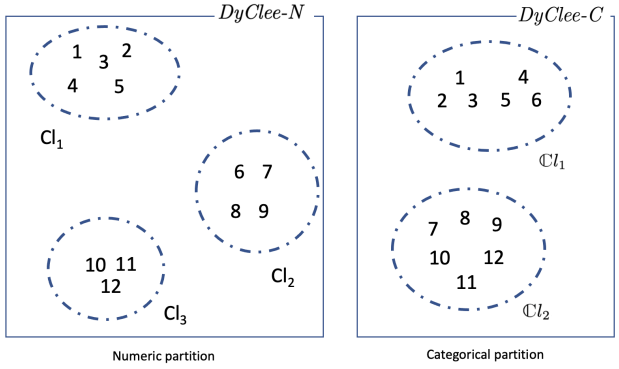

Fig. 5. Example of numeric and categorical sample clusters

The product of partitions is applied to P_s and $\mathbb{P}_s$.

Figure 6 shows the final mixed partition, standing for the final mixed clustering, given by $\mathcal{P} = \{\mathcal{C}l_1, \mathcal{C}l_2, \mathcal{C}l_3\}$ with $\mathcal{C}l_1 = \{1, 2, 3, 4, 5\}$, $\mathcal{C}l_2 = \{7, 8, 9\}$ and $\mathcal{C}l_3 = \{10, 11, 12\}$. Sample 6 does not appear in the mixed clusters because it appears alone in a cluster which is considered *outlier*.

3.4 DyClee-N&C *parameters*

*DyClee-N&C*supports several optional parameters that allow the user to improve clustering results by adding some knowledge to the data. These parameters are tuned according to the problem at hand and they may be evaluated with cluster validity methods (Liu et al., 2010). Some of these parameters are listed below.

Parameter forget_method – Different forgetting processes can be used to achieve correct tracking of the data evolution process. The default is *forget_method*=NONE, for which no forgetting process is applied. The other options apply different forgetting strategies.

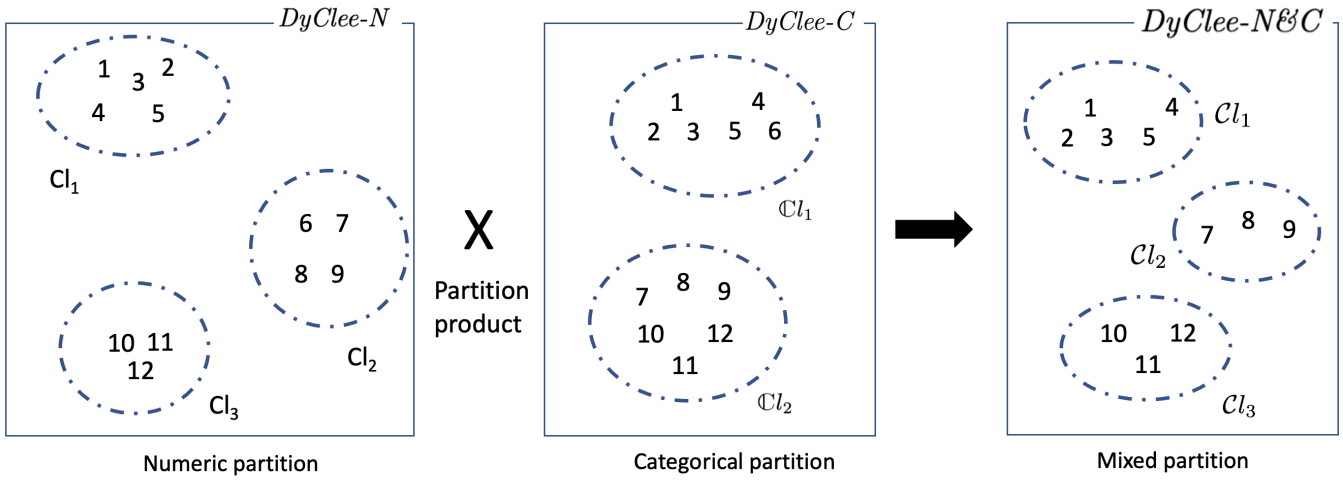

Fig. 6. Final partition $\mathcal{P}$ obtained with *DyClee-N&C*

Parameter Unclassed_accepted – Final clusters are composed of dense (body of the cluster) and semi-dense (edge of the cluster) μ-clusters. Outliers are considered unrepresentative or noise and their samples are rejected. Depending on the context, it may be interesting to assign all the data to some cluster and not reject outliers. *Unclassed_accepted*=OFF means that all the samples must be assigned to a cluster, i.e. there is no outlier rejection. Default mode is *Unclassed_accepted*=ON.

Parameter minimum_mc – This parameter sets the minimum number of samples that a cluster must contain to be considered as a final cluster. Default mode is *minimum_mc*=FALSE. *minimum_mc*=TRUE should be used when the focus of the analysis is on populated cluster profiles.

Parameter multi-density – This parameter decides about using the global or the local density threshold. *multi − density*=TRUE means that local density analysis is required. Default mode is *multi − density*=FALSE.

Parameter n_clusters – This parameter allows to retain the k most important clusters as final clusters. Samples belonging to the remaining clusters are re-assigned to one of the k retained clusters. This requires *n_clusters*=ON. default mode is *n_clusters*=OFF.

4. EVALUATION

This section presents the evaluation of *DyClee-N&C* on the public data set *Automobile* (Schlimmer, 1987) with a comparison with two other mixed clustering approaches *K-Prototype* (Huang, 1998) and *CAH mixte* (Ward Jr, 1963). This data set has been selected due to its characteristics i.e 205 samples, with 26 attributes and 6 clusters. A sample corresponds to one vehicle and the features correspond to automotive components (number of doors, number of cylinders, ...) and to information about the vehicle in a more general way (vehicle length, weight, ...). The classes in this data set correspond to the risk rating for the vehicle. More the value is high more the vehicle presents a risk. The percentage of numeric features is 60 percent, while the percentage of categorical ones is 40 percent. The categorical features are composed of nominal terms.

The evaluation of the formed clusters is based on the internal silhouette index ((Rousseeuw, 1987)). Applying an internal validation index allows to evaluate if the obtained cluster partitions are valid from an agnostic point of view i.e low internal cluster inertia within each cluster (cluster compactness) and high inter cluster inertia between all clusters (cluster separability). Generally, the euclidean distance is used to evaluate the different proximities needed to compute the silhouette but here to apprehend mixed attributes the *Gower* distance is used. (Gower, 1971), $D_g(x_i, x_k)$, given by the equation (2) is used to measure the dissimilarity between two samples.

$$D_g(x_i, x_k) = 1 - \frac{1}{d}\sum_{j=1}^{d} s_{ik}^j \tag{2}$$

with d the number of features and s_{ik}^j the similarity between samples x_i and x_k for the j^{th} feature. The similarity index s_{ik}^j is evaluated according to the feature type:

- categorical feature: $s_{ik}^j = 1$ if $x_i^j = x_k^j$, else 0.
- numeric feature: $s_{ik}^j = 1$ - $\frac{|x_i^j - x_k^j|}{R_j}$ with R_j the maximal difference for the j^{th} feature.

In mixed *K-Prototype* and *CAH*, the number of clusters must be specified in input. As this type of information is not supposed to be known a priori, we varied the value of the number of clusters k so that $2 \leq k \leq \sqrt{n}$ with n the number of samples in the data set. The limiting value $\sqrt{n}$ is chosen in order to avoid the situation where the number of clusters tends to be equal to the number of samples. In this configuration, the number of samples assigned to the clusters is low and the similarity between the individuals composing them is high. The silhouette score associated with each sample then becomes abnormally high, which biases the interpretation of the evaluation of the obtained clusters. For the necessary parameters related to *DyClee* mixed, the value given to the size of the numerical μ-

Methods	*DyClee-N&C*	K-Prototype	CAH mixed
S-score	**0.23**	0.17	0.22
N° of clusters	**2**	14	14

Table 1. Silhouette scores

clusters noted ϕ varies between 0.05 and 0.5 with a step of 0.05. Concerning the threshold of similarity of *Hamming* noted *Seuil_Ham*, used at the time of the phase of assignment of a categorical sample to the categorical μ-cluster, its value varies between 0 and 1 with a step of 0.1.

The clustering algorithms mixed *K-Prototype* and *CAH* consider all samples as representative (*i.e* no outlier). Therefore, the *unclassed_accepted* parameter of *DyClee-N&C* is set to OFF in order to keep all μ-clusters formed by *DyClee-N&C* for this part of the experiments. The values of the parameters *minimum_mc* and *n_clusters* are tested in order to refine the final clusters. These parameters are first deactivated to evaluate the clusters formed by *DyClee-N* and *DyClee-C* which correspond respectively to the clusters found from *KD-Tree* and *LSH*. In this experiment, the size of the numerical μ-clusters is ϕ=0.3, the similarity threshold to be satisfied for a categorical sample to be assigned to a categorical μ-cluster is 0.5.

Table 1 gives the silhouette scores (noted S-score) for the clusters formed by *DyClee-N&C*, *K-Prototype*, and mixed *CAH* methods along with the associated number of clusters. *DyClee-N&C* obtains a much better silhouette score than the other two methods.

The number of clusters retained to obtain this silhouette score is 2 for *DyClee-N&C* and 14 for the other methods. *DyClee-N&C* finds a number of clusters that is closest to the number of classes present in the *Automobile* data set.

Complexity– The complexity of *DyClee-N&C*can be determined as the maximum of the complexities of *DyClee-N*, *DyClee-C*, and of the product of partitions. After determining these complexities, it has been shown that the complexity of *DyClee-N&C*is polynomial.

5. CONCLUSIONS AND PERSPECTIVES

This paper presents *DyClee-N&C*, the mixed version of the original *DyClee*, named *DyClee-N* in this paper, algorithm of which a purely categorical version, named *DyClee-C* had already been proposed. *DyClee-N&C* subsumes both the numeric and categorical feature based algorithms *DyClee-N* and *DyClee-C*. *DyClee-N&C* hence represents a versatile and usable dynamic clustering algorithm for data described by all types of features. This gives it a definite advantage for some situation assessment and diagnosis applications in which situations are characterized by numeric and categorical features. This is exemplified with a data set used to assess risk in the automobile domain.

Let us notice that hierarchical clustering versions of *DyClee-N* and *DyClee-C* have been implemented. Although these only work for "static" clustering, they may be interesting to decide at which aggregation level the clustering must be considered to match business knowledge. Future work will consider to produce a hierarchical clustering version of *DyClee-N&C*.

REFERENCES

Barbosa Roa, N., Travé-Massuyès, L., and Grisales, V.H. (2019). Dyclee: Dynamic clustering for tracking evolving environments. *Pattern Recognition*. doi: doi.org/10.1016/j.patcog.2019.05.024.

Gionis, A., Indyk, P., Motwani, R., et al. (1999). Similarity search in high dimensions via hashing. In *Vldb*, volume 99, 518–529.

Gower, J.C. (1971). A general coefficient of similarity and some of its properties. *Biometrics*, 27(4), 857–871.

He, Z., Xu, X., and Deng, S. (2005). Clustering mixed numeric and categorical data: A cluster ensemble approach. *arXiv preprint cs/0509011*.

Huang, Z. (1998). Extensions to the k-means algorithm for clustering large data sets with categorical values. *Data Mining and Knowledge Discovery 2*, 283–304.

Hummel, M., Edelmann, D., and Kopp-Schneider, A. (2017). Clumix: Clustering and visualization of mixed-type data. *URL: https://pdfs. semanticscholar. org/1e65/755051c4b749fac17a23ff93924157acacdd. pdf*.

Indyk, P. and Motwani, R. (1998). Approximate nearest neighbors: towards removing the curse of dimensionality. In *Proceedings of the thirtieth annual ACM symposium on Theory of computing*, 604–613. ACM.

Ji, J., Pang, W., Zhou, C., Han, X., and Wang, Z. (2012). A fuzzy k-prototype clustering algorithm for mixed numeric and categorical data. *Knowledge-Based Systems*, 30, 129–135.

Leskovec, J., Rajaraman, A., and Ullman, J.D. (2014). *Mining of massive datasets*. Cambridge university press.

Liu, Y., Li, Z., Xiong, H., Gao, X., and Wu, J. (2010). Understanding of internal clustering validation measures. In *2010 IEEE International Conference on Data Mining*, 911–916. IEEE.

Maneewongvatana, S. and Mount, D.M. (1999). On the efficiency on nearest neighbor searching with data clustered in lower dimensions. In *ICCS '01 Proceedings of the International Conference on Computational Sciences-Part I*.

McParland, D. and Gormley, I.C. (2016). Model based clustering for mixed data: clustmd. *Advances in Data Analysis and Classification*, 10(2), 155–169.

Obry, T., Travé-Massuyès, L., and Subias, A. (2019). Dyclee-c: a clustering algorithm for categorical data based diagnosis. In *DX'19–30th International Workshop on Principles of Diagnosis*.

Rousseeuw, P.J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and applied Mathematics*.

Schlimmer, J.C. (1987). UCI machine learning repository. URL `http://archive.ics.uci.edu/ml`.

Ward Jr, J.H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, 58(301), 236–244.