COPILS: COmParIson of Linguistic Summaries

Grégory Smits[1] and Marie-Jeanne Lesot[2]

[1] IMT Atlantique, Brest, France
Gregory.Smits@imt-atlantique.fr
[2] Sorbonne Université CNRS, LIP6, Paris, France
Marie-Jeanne.Lesot@lip6.fr

Abstract. When tabular data cannot be directly mined, due to their large size or for privacy reasons, their summary may still be available for analysis. Fuzzy linguistic summaries are composed of sentences describing a multivariate data distribution using, possibly personalized, terms taken from a fuzzy vocabulary. This paper introduces an algorithmic solution to the comparison of summaries so as to provide users with a linguistic description of the data changes between datasets to be compared. A first strategy processes exhaustive summaries containing one sentence for each of the subspaces that can be formed using terms from the vocabulary. A second strategy is proposed for condensed summaries, that involve informative sentences only. Experimentation conducted on artificial datasets confirm the relevance of this second strategy in terms of computational cost and data changes that can be tracked.

1 Introduction

Fuzzy linguistic summaries offer legible overviews on the content of tabular data sets, through a set of sentences that follow predefined syntactic protoform, e.g. $Q\,\mathcal{X}\,are\,P$ (see e.g. [4]). The instantiation of this schema can for instance lead to *some data are* A_1*.medium and* A_2*.low*, where A_1 and A_2 are descriptive features of a data set $\mathcal{X}$. Such summaries provide concise and personalised insights into the data content: the conciseness is due to the linguistic descriptions of the clusters that compose the data, together with some rarer behaviours; the personalisation property comes from the use of linguistic terms, such as *medium* or *low*, that can be defined individually for each user.

This paper addresses the task of comparing linguistic summaries extracted from two data sets $\mathcal{X}$ and $\mathcal{X}'$, described by the same features and linguistic terms: $\mathcal{X}$ and $\mathcal{X}'$ may correspond to two subpopulations of a data set, e.g. students with different majors, or to two data sets collected at different time steps, e.g. students of two different years. The aim is to provide insights on the data distribution differences through their characterisation by linguistic summaries.

The question of comparing two data sets, or chronological updates of a data set, has been largely studied by the data mining and machine learning communities, in the tasks of data drift and concept drift detection and processing (see e.g. [2] for a general survey). Yet, these methods usually consider that the data

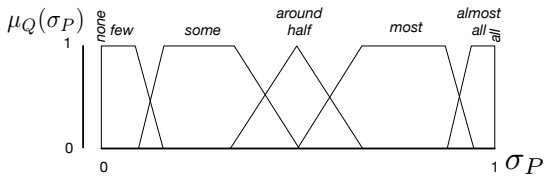


Fig. 1. Partition $\mathcal{Q}$ defining 7 quantifiers to describe relative cardinalities

are available. Now it may be the case that the data cannot be shared, due to privacy or storage constraints. In such cases, data mining tasks can be applied to their linguistic summaries instead of the data themselves [9]. For such a case, the paper proposes COPILS, a method for the COmParIson of fuzzy Linguistic Summaries: it generates *differential*, or *comparative*, *summaries* that allow to compare data sets in a legible way, when the whole data is not accessible: it generates data explanations intelligible to users, in the dynamic eXplainable Artificial Intelligence (XAI) framework.

The paper is structured as follows: Section 2 describes formally the considered context of fuzzy linguistic summaries and existing works for their comparison. Section 3 presents the proposed COPILS approach, that is experimentally studied on synthetic data sets in Section 4. Section 5 concludes the paper.

2 Background and Context

2.1 Formal Definition of Fuzzy Linguistic Summaries

$\mathcal{X}$ denotes a data set of n points, $\mathcal{X} = \{x_1, \ldots, x_n\}$ described by d features, that can be numerical or categorical, $\{A_1, \ldots, A_d\}$, respectively defined on domain D_j, $j = 1 \ldots d$. Each data point is denoted $x = (A_1.x, \ldots, A_d.x)$. A fuzzy vocabulary $\mathcal{V}$ is defined as a set of fuzzy partitions $\mathcal{V} = \{\mathcal{V}_1, \ldots, \mathcal{V}_d\}$, $\mathcal{V}_j$ is a triple that associates feature A_j with modalities, defined as fuzzy sets on its domain D_j, and with linguistic labels: $\langle A_j, \{\mu_{j1}, \ldots, \mu_{jq_j}\}, \{l_{j1}, \ldots, l_{jq_j}\}\rangle$ (see Figure 2 for an illustrative example). It is assumed that the linguistic variables discretizing an attribute domain, form a strong partition: $\forall j \in \{1..d\}, \forall y \in D_j$, $\sum_{s=1}^{q_j} \mu_{js}(y) = 1$. In addition, as illustrated in Figure 1, a partition $\mathcal{Q}$ is defined on the universe $[0, 1]$ to describe relative cardinalities.

Sentences Each sentence in a fuzzy linguistic summary is an instantiation of a predefined scheme, called protoform, written $Q\,\mathcal{X}\,are\,P$ or $Q\,R\,\mathcal{X}\,are\,P$ [4]: Q is the linguistic label of a fuzzy quantifier, taken from $\mathcal{Q}$. P, called the summarizer, and R, called the qualifier, are defined as conjunctions of terms taken from the vocabulary $\mathcal{V}$. This paper focuses on the first type, as the second one can be seen as an instantiation thereof, applied to a fuzzy data set defined as the extraction,

through the fuzzy filter R, of $\mathcal{X}$. For the data represented in Figure 2, this protoform can be illustrated by the sentence *some data are* A_1*.high and* A_2*.low*.

Each sentence is associated with a degree of truth that measures the extent to which it is adequate to represent the data [4]: $\tau(Q\,\mathcal{X}\,are\,P) = \mu_Q(\sigma_P(\mathcal{X}))$, with $\sigma_P(\mathcal{X}) = \frac{1}{|\mathcal{X}|}\sum_{x\in\mathcal{X}}\mu_P(x)$. If $P = A_{(1)}.m_{(1)}$ and $\ldots$ $A_{(k)}.m_{(k)}$ where $A_{(i)}$ are features and $m_{(i)}$ linguistic labels of modalities taken from their associated partitions, with respective membership functions $\mu_{(i)}$, its membership function μ_P is defined as $\mu_P(x) = \top_{i=1}^{k}\mu_{(i)}(A_{(i)}.x)$, where $\top$ is a t-norm.

Summaries A summary $\mathcal{S}$ is then a set of sentences $Q\,\mathcal{X}\,are\,P$ describing the data set. *Exhaustive summaries* contain one sentence for each possible P, i.e. for each possible combination of $A.m$, including the absence of feature A. For each of them, the most appropriate quantifier, i.e. $Q \in \mathcal{Q}$ with maximal $\mu_Q(\sigma_P(\mathcal{X}))$, is selected. Such summaries have a length exponential in the number of features: their number of sentences equals $\Pi_{j=1}^{d}(1+q_j)$. An obvious filtering step excludes sentences containing summarizers P for which the selected quantifier Q is *none*, it still leads to a summary with exponential length.

Several approaches have been proposed to prune the set of sentences to the most relevant ones, leading to *condensed summaries*. Some of them do not assess sentences individually, but globally, quantifying their redundancies, e.g. exploiting known relations between $\tau(Q\,\mathcal{X}\,are\,P)$ and $\tau(Q\,\mathcal{X}\,are\,P\,and\,P')$. These methods differ by their pruning criteria [8, 12]. Others avoid the generation of non-relevant summaries that are later discarded, using integrated approaches: some exploit the relations between linguistic summaries and association rules [5, 6], others exploit the same principle of anti-monotonicity of quality criteria, e.g. considering the degree of focus, in addition to the truth degree [11, 13].

2.2 Distances at the Fuzzy Linguistic Level

The central question addressed in this paper is the comparison of data sets considered through the lens of a fuzzy vocabulary, that can be seen as a data rewriting tool. Distance measures at different levels can then be considered.

First for values corresponding to a single feature, $y, y' \in D$, a distance can be computed taking into account the associated fuzzy partition V, more precisely their membership degrees to the modalities best associated to them, as well as the number of modalities that separate them [3]:

$$d_V(y, y') = \frac{1}{q_V - 1} \times |\mu_{I(y)}(y) - \mu_{I(y')}(y') + I(y') - I(y)|, \tag{1}$$

where q_V is the number of modalities in V and $I(y) \in \{1, \ldots, q_V\}$ the index of the area in which y falls, defined by the lower bounds of the modality cores.

For sentences, it has been proposed in [14] to compare $s = Q\,\mathcal{X}\,are\,P$ and $s' = Q'\,\mathcal{X}'\,are\,P'$, with respective truth values τ and τ', through the distance

$$d(s, s') = 1 - \min(sim(Q, Q'), sim(P, P'), sim(\tau, \tau')). \tag{2}$$

The use of the minimum implies that the three similarity components all play the same role in the comparison. This distance has been used to structure the possibly large flat set of sentences into groups of similar sentences [14], but it does not appear appropriate for the comparative task considered in this paper. Indeed, sentences that contain the same summarizer with different quantifiers must be considered as closer than sentences that apply the same quantifier to different summarizers. Section 3.2 proposes a distance with these semantics.

Besides, the question of comparing linguistic summaries can be considered as related to that of emerging patterns [1, 10]: the latter characterize a data set by contrast to a reference one, through frequent itemsets whose quality is measured as the quotient of their supports computed on the two data sets.

3 Proposed Approach: COPILS

This section describes COPILS, which stands for COmParIson of Linguistic Summaries, that allows to compare the fuzzy linguistic summaries extracted from two tabular data sets, with the aim to allow a user to get a legible insight on their differences. It considers in turn the cases of exhaustive and condensed summaries.

3.1 Exhaustive Summary Comparison

The comparison of exhaustive summaries is an easy task, as any summarizer P appears in exactly one sentence in each summary, except if it is associated with the quantifier *none*. For a given P, only three types of differences can be observed:

- P is absent of one of the summaries, corresponding to a case of data addition or removal in the fuzzy subspace described by P.
- the quantifiers selected during the summarization step differ, i.e. the two sentences are of the form $s = Q\,\mathcal{X}\,are\,P$ and $s' = Q'\,\mathcal{X}'\,are\,P$, $Q \neq Q'$. This case corresponds to a cardinality change in the subspace described by P. The previous case is a special case of this one, when Q xor Q' equals *none*.
- the quantifiers are identical, only the truth degrees differ.

COPILS extracts the couples of sentences s and s' with the same summarizer but different quantifiers. For each of them, it generates a differential characterisation of the following form (see illustrations in the next subsection):

- if $Q = none$ and $Q' \neq none$, the apparition of a point group in an initially empty subspace is characterised, in the differential summary, as
  $ADDED\ dist. = d(none, Q')$: $Q'\ \mathcal{X}\ are\ P\ truth{=}\mu_{Q'}(\sigma_P(\mathcal{X}'))$
  where d is the quantifier distance $dist_{quant}$ defined in Equation (3) below,
- if $Q \neq none$ and $Q' = none$, the disappearance of a point group is characterised as
  $REMOVED\ dist. = d(Q, none)$: $Q\ \mathcal{X}\ are\ P\ truth{=}\mu_Q(\sigma_P(\mathcal{X}))$,
- if $Q \neq none$ and $Q' \neq none$, the cardinality change is characterised as
  $MODIFIED\ dist. = d(Q, Q')$: $Q\ \mathcal{X}\ are\ P \quad truth{=}\mu_Q(\sigma_P(\mathcal{X}))$
  $\Rightarrow Q'\ \mathcal{X}'\ are\ P\ truth{=}\mu'_Q(\sigma_P(\mathcal{X}'))$.

For the $d(Q, Q')$ distance between quantifiers, we use the following definition

$$dist_{quant}(Q, Q') = \frac{|I(Q) - I(Q')|}{|\mathcal{Q}|-1}, \tag{3}$$

where $I(Q)$ (resp. $I(Q')$) is the index of quantifier Q (resp. Q') in the $\mathcal{Q}$ partition of size $|\mathcal{Q}|$. This way, using e.g. the partition shown in Fig. 1, a modification from *most* to *some*, associated to the distance |3-5|/(7-1) = 1/3, is more important than from *most* to *around half*, associated to |5-4|/(7-1) = 1/6. COPILS displays the data changes in decreasing order of $dist_{quant}(Q, Q')$, so as to prioritize the most significant ones, increasing the legibility of the generated summaries.

Properties A first specificity of the proposed COPILS approach is that a summarizer P associated with the same quantifier in both summaries does not lead to a comparative sentence, even if the truth degree varies: we consider that a modification that is not important enough to induce a quantifier change does not deserve to be mentioned. This means that differences are measured up to the linguistic terms used in the partition $\mathcal{Q}$ associated to the relative cardinalities. This is a way to integrate the user in the result expression, taking into account the granularity that he/she indicates as making sense to him/her.

A second property is that the comparison is as exhaustive as the initial summaries are: a quantifier modification for a complex summarizer P that is made of a conjunction of several modalities implies that at least one of its components is also associated to a quantifier modification. As a consequence, they all generate a sentence in the output differential summaries, possibly leading to long results. It may be relevant to select some of them, applying some pruning strategy, in the same sense as the ones discussed in Section 2.1. However it is difficult to determine whether the user wishes to focus on the differences on the maximal summarizers, or, on the contrary, on the minimal ones in terms of number of modalities. Thus no pruning is performed in the current version of COPILS.

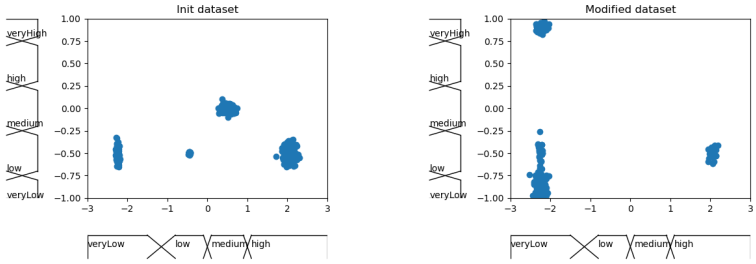
Globally, the sentence distance on which COPILS relies can be formalised as follows: if the summarizers differ, it is considered to be arbitrarily high; otherwise, the ordering strategy compares the quantifiers, i.e.

$$dist_{sent}(s, s') = \begin{cases} 1 & \text{if } P \neq P' \\ dist_{quant}(Q, Q') & \text{otherwise.} \end{cases} \tag{4}$$

Contrary to the distance reminded in Eq. (2) [14], it gives an asymmetrical role to the summarizer and quantifier and it does not depend on the truth degrees.

Illustrative Example Figure 2 shows an illustrative 2D initial data set $\mathcal{X}$ (left) and its modified version $\mathcal{X}'$ (right), together with the fuzzy partitions associated to A_1 (x-axis) and A_2 (y-axis), as well as the exhaustive and condensed (see definition in the next subsection) summaries of $\mathcal{X}$. The four clusters composing $\mathcal{X}$ are modified as follows to define $\mathcal{X}'$:

– the cardinality of the cluster *'A_1.veryLow and A_2.low'* goes from 50 to 25,



Exhaustive summary of $\mathcal{X}$:10 sentences
'few data are A_1.veryLow' τ=1 *'few data are A_1.veryLow and A_2.low'* τ=1
'some data are A_1.low' τ=1 *'some data are A_1.low and A_2.low'* τ=1
'some data are A_1.medium' τ=1 *'some data are A_1.medium and A_2.medium'* τ=1
'some data are A_1.high' τ=1 *'some data are A_1.high and A_2.low'* τ=1
'most data are A_2.low' τ=1
'some data are A_2.medium' τ=1
Condensed summary of $\mathcal{X}$: 5 sentences
'most data are A_2.low' τ=1 *'few data are A_1.veryLow and A_2.low'* τ=1
'some data are A_1.low and A_2.low' τ=1
'some data are A_1.medium and A_2.medium' τ=1
'some data are A_1.high and A_2.low' τ=1

Fig. 2. Illustrative case: (left) initial dataset $\mathcal{X}$, (right) its modified version $\mathcal{X}'$, (bottom) fuzzy linguistic summaries, exhaustive and condensed, describing $\mathcal{X}$.

- the cardinality of the cluster *'A_1.high and A_2.low'* goes from 200 to 20,
- the cluster initially located at the intersection of *'A_1.medium'* and *'A_2.medium'* moves to *'A_1.veryLow and A_2.veryLow'*,
- the cluster initially located at the intersection of *'A_1.low'* and *'A_2.low'* moves to *'A_1.veryLow and A_2.veryHigh'*.

After discarding the sentences associated with the *'none'* quantifier, the exhaustive summary of $\mathcal{X}$, that may contain up to $(q_1 + 1) \times (q_2 + 1) = 30$ sentences, contains the 10 sentences shown below the graphical representation (Fig. 2).

COPILS then leads to the following differential summary, with length 12:

- MODIFIED *dist.* = 2/3: *'few data are A_1.veryLow' truth=1*
  $\Rightarrow$ *'almost all data are A_1.veryLow' truth=1*,
- ADDED *dist.* = 0.5: *'around half data are A_2.veryLow' truth=0.69*,
- ADDED *dist.* = 0.5: *'around half data are A_1.veryLow and A_2.veryLow' truth=0.69*,
- REMOVED *dist.* = 1/3: *'some data are A_1.low' truth=1*,
- REMOVED *dist.* = 1/3: *'some data are A_1.medium' truth=1*,
- REMOVED *dist.* = 1/3: *'some data are A_1.low and A_2.low' truth=1*,
- REMOVED *dist.* = 1/3: *'some data are A_1.medium and A_2.medium' truth=1*,
- MODIFIED *dist.* = 1/3: *'most data are A_2.low' (truth=1)*
  $\Rightarrow$ *'some data are A_2.low' truth=0.9*,
- ADDED *dist.* = 1/6: *'few data are A_1.veryLow and A_2.medium' truth=1*,

- MODIFIED *dist.* = 1/6: *'some data are* A_1*.high' truth=1*
  $\Rightarrow$ *'few data are* A_1*.high' truth=1*,
- MODIFIED *dist.* = 1/6: *'some data are* A_2*.medium' truth=1*
  $\Rightarrow$ *'few data are* A_2*.medium' truth=1*,
- MODIFIED *dist.* = 1/6: *'some data are* A_1*.high and* A_2*.low' truth=1*
  $\Rightarrow$ *'few data are* A_1*.high and* A_2*.low' truth=1.*

It exhibits all the differences and so covers all the expected changes. Yet it may be seen as still difficult to read, because of redundancies.

3.2 Condensed Summary Comparison

The previous comparison of exhaustive summaries generates a high number of differential sentences, at least as many as the longest one. This section presents the COPILS variant for comparing condensed summaries, obtained after pruning redundancies in exhaustive ones, so as to generate shorter differential summaries.

Considered Condensed Summary Definition Among existing pruning strategies (see the brief discussion in Section 2.1), we consider the approach relying on maximal sentences [13]. The latter are defined as instanciations of the protoform $s = Q\,\mathcal{X}\,are\,P$ such that no instanciation of $s' = Q\,\mathcal{X}\,are\,P and\,P'$ holds: a maximal sentence prunes the sentences covered by its summarizer if they are attached to the same quantifier, even if their truth values vary. For instance, knowing that *'some data are* A_1*.medium and* A_2*.medium* holds, sentences for A_1*.medium* and A_2*.medium* individually are kept only if they cover a different ratio of the data set. A condensed summary is then defined as a summary that only contains maximal sentences, the other ones being removed, leading to a significantly lower length as compared to the exhaustive summary (see bottom part of Fig. 2 for an illustrative example and Section 4.2 for a quantification).

Such condensed summaries obviously overcome the main limitation of exhaustive summaries which comes from their huge number of sentences. However, a sentence with a summarizer P may have no counterpart in the other one, as it may have been pruned. The comparison of condensed summaries thus requires a matching step, to pair their respective sentences.

Optimal Matching Step In order to perform this matching step, the COPILS approach we propose relies on a variant of the stable marriage algorithm: two sentences are paired if no closer pair can be formed with any other unpaired sentence. Yet, different from the initial stable marriage problem, the numbers of sentences are not necessarily equal. Thus, some sentences can remain unmatched.

To perform pairing, it is necessary to define a distance between summarizers. Then only pairs of sentences for which the latter is acceptable (i.e. lower than a user-defined threshold η) are kept. We propose to define the distance between summarizers, defined as sets of conjuncts, as:

$$d_{sum}(P, P') = \frac{1}{\max(|P|, |P'|)} \sum_{j=1}^{d} d_{mod}^{j}(P, P'), \tag{5}$$

where $d^j_{mod}(P, P')$ compares the conjuncts in P and P' that apply to feature j. If P xor P does not contain any, it equals 1; otherwise, it depends on the absolute difference of the indices of these modalities in $\mathcal{V}_j$, in the same spirit as the quantifier distance from Eq. (3), replacing the quantifier partition with $\mathcal{V}_j$.

The distance between two sentences in condensed summaries is defined as a prioritised aggregation of the distances at the summarizer and quantifier levels:

$$dist_{sent} = (s, s') = \begin{cases} 1 & \text{if } d_{sum}(P, P') > \eta, \\ \max(d_{sum}(P, P'), d_{quant}(Q, Q')) & \text{otherwise.} \end{cases} \quad (6)$$

As compared to the sentence distance reminded in Eq. (2) [14], this distance shares the same characteristics as the exhaustive summaries distance defined in Eq. (4): it does not depend on the truth degrees and it gives an asymmetrical role to the summarizer and quantifier comparison, combining them in a hierarchical process. Indeed, the summarizer comparison plays a prominent role, the quantifier distance is only involved in a second step.

Form of the Generated Differential Summary The COPILS variant for condensed summaries yields four types of differential characterisations, adding one to the exhaustive case. Situations of data addition (resp. removal) correspond to cases when a sentence from the modified (resp. initial) summary has not been paired, because no close enough summarizer has been found. Paired sentences are decomposed into two cases, depending on whether their summarizers are identical, interpreted as modifications, as previously, or not, interpreted as possible data moves from a subspace to another, leading to

POSSIBLE MOVE $dist. = d(s, s')$: $Q\mathcal{X}$ *are* P $\quad$ *truth*=$\mu_Q(\sigma_P(\mathcal{X}))$
$\Rightarrow$ $Q'\mathcal{X}'$ *are* P' *truth*=$\mu'_Q(\sigma'_P(\mathcal{X}'))$.

Again, the differential characterisations are displayed by COPILS in decreasing order of their attached distance. The latter is defined as the distance between the two paired sentences in case of a POSSIBLE MOVE, and the distance between the quantifiers of the sentences in the three other cases.

Illustrative Example Considering the data shown on Fig. 2, the condensed summary of $\mathcal{X}$ only contains five sentences (see bottom part of the figure): for instance the cluster with the highest values on feature A_2 is described by a single sentence: *'some data are A_1.medium and A_2.medium'*. Indeed, the property of maximal sentences guarantees that the sentences *'some data are A_1.medium'* and *'some data are A_2.medium'* also hold and they are not included. COPILS then generates the following output of length 6:

- POSSIBLE MOVE $dist. = \max(1/6, 7/12)$: *'some data are A_1.low and A_2.low', truth=1* $\Rightarrow$ *'around half data are A_1.veryLow and A_2.veryLow' truth=0.69*,
- POSSIBLE MOVE $dist. = \max(1/6, 1/3)$: *'some data are A_1.medium and A_2.medium', truth=1* $\Rightarrow$ *'few data are A_1.veryLow and A_2.medium' truth=1*,
- MODIFIED $dist. = 1/3$: *'most data are A_2.low' truth=1*
$\Rightarrow$ *'some data are A_2.low' truth=0.9*.

- ADDED *dist.* = 1/3: *'some data are A_1.veryLow and A_2.veryHigh' truth=1*,
- ADDED *dist.* = 1/6: *'almost all data are A_1.veryLow' truth=1*,
- MODIFIED *dist.* = 1/6: *'some data are A_1.high and A_2.low' (truth=1)*
  $\Rightarrow$ *'few data are A_1.high and A_2.low' truth=1.*

This result also covers the expected changes between $\mathcal{X}$ and $\mathcal{X}'$, in a more legible result than in the exhaustive case, pointing out more easily the main changes. Moreover, allowing partial matching between the compared summarizers when pairing sentences makes it possible to suggest possible data moves.

4 Experimental Results

This section studies the cost of the differential summary generation performed by the two COPILS variants, i.e. for exhaustive vs condensed summaries, both in terms of summary lengths and of computation time, on synthetic data.

4.1 Synthetic Data Generation

The use of synthetic data allows to control the summarisation task complexity, by selecting the generation parameters, as detailed below.

To obtain meaningful summaries, and subsequently meaningful comparisons, a crucial issue is to ensure an adequacy between the discretization of the attribute domains operated by the fuzzy partitions and the distribution of the data [7]. To achieve this aim, we propose a generation process that takes as input the fuzzy vocabulary. More precisely, we consider up to 8 numerical features with different domain ranges and a fuzzy partition manually defined for each of them, with respective numbers of modalities varying from three to six. Compact and ellipsoidal clusters are then generated by multivariate Gaussian distributions. Their means and standard deviation are randomly chosen in bounded ranges, so as to ensure that each cluster is mostly covered by one modality on each feature, ensuring their adequacy wrt the considered vocabulary.

In a second step, given an initial data set $\mathcal{X}$ generated using this process, we produce a modified version $\mathcal{X}'$ in such a way that the modifications that should appear in the comparative summary are known in advance. To do so, each cluster in $\mathcal{X}$ is modified with a uniform probability, with two types of modifications allowed: (i) a cardinality modification randomly removes some of its members or generates new members according to the Gaussian probability distribution it is associated with; (ii) a subspace modification removes the considered cluster and creates a new one in another randomly chosen subspace. This subspace may vary from the initial one on a single dimension or all of them.

The hyper-parameters of the data generation process are the fuzzy vocabulary, the number of features to consider and the number of clusters to generate. The other parameters are set randomly. Note that the total number of data points is not varied: it only influences the preliminary step of summarising $\mathcal{X}$ and $\mathcal{X}'$. COPILS takes as an input the resulting summaries, not depending on

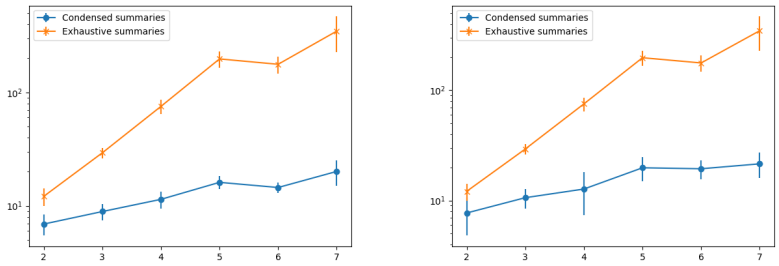


Fig. 3. Length (in log-scale), of (left) $\mathcal{X}$ summaries and (right) $\mathcal{X}$ and $\mathcal{X}'$ comparative summaries, in the exhaustive and condensed cases, wrt the number of features.

the data set points. The data shown in Fig. 2 has been generated by this procedure, for 2 features, respectively associated with partitions of 4 and 5 modalities, and 4 clusters.

4.2 Differential Summary Length

The first criterion we consider to assess the differential summaries COPILS generates is the number of characterisations they contain, i.e. their length.

Input Summary Length As a preliminary observation, the left part of Figure 3 displays the average and standard deviation of $\mathcal{X}$ summary length for increasing number of features, computed on 10 runs for each setting. The slight decrease observed with 6 features is due to the fact that, for this setting, the randomly generated datasets were very sparse. It shows that, as expected, for more than 4 features, exhaustive summaries contain more than one hundred sentences, making them intractable for the end user. The size of condensed summaries also increases exponentially in the number of features, but with a much lower power, offering more legible results.

Output Summary Length The right part of Figure 3 shows the average and standard deviation of the COPILS differential summary lengths, for the same data sets. As expected, for the exhaustive case, they are identical to the lengths of $\mathcal{X}$ summaries; for the condensed case, they can be slightly higher: they also are intractable above three features for the exhaustive case, and acceptable for the condensed case, offering a preliminary validation of this latter approach.

In order to study in more details the parameters that influence the differential summary length, we investigate the effect of the number of clusters and the associated number of modifications: setting the number of features to 4, the data generation process described in Section 4.1 is applied for various numbers of clusters, which correlates with the number of modifications between $\mathcal{X}$ and $\mathcal{X}'$.

Table 1. Length of comparative summaries for various number of clusters and data modifications

Configuration	Nb. of clusters	2	3	4	5
	Avg data changes	1.4	2.2	2.9	3.9
Summary Length	Exhaustive	41.4	51.36	66.73	75.6
	Condensed	6.7	8.5	11.6	12

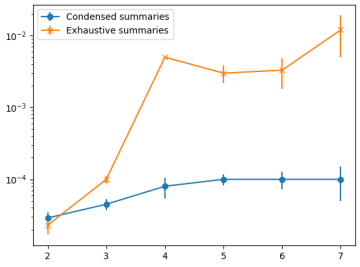


Fig. 4. Computational cost (in seconds, log-scale) for the comparative summaries generation in the exhaustive and condensed cases wrt the number of features.

Table 1 displays the length of the comparative summaries, averaged on 20 runs, together with the average number of modifications. It shows that the number of clusters, and consequently the number of modifications present in the data, also influences the lengths. Again the result length is intractable in the case of exhaustive summaries, while remaining reasonable for the condensed one.

4.3 Computation Time

Figure 4 shows that, as expected, the computation time required to output the differential summaries has an exponential dependency on the number of features, with a higher degree for exhaustive than for condensed summaries. This is a consequence of the size of the input to be processed.

It can thus be observed that, for 3 features or more, the increase in computation time induced by the matching step is negligible as compared to the gain of processing a much smaller input: the reduced number of sentences to be processed compensates for the matching overhead, except for 2 features. This very special case cannot be seen as restricting the relevance of the proposed comparison of condensed summaries.

5 Conclusion and Future Works

In order to provide a legible view of the changes between two datasets with the same descriptive space, when only their summaries are available, COPILS exhibits additions, removals and possible moves in the data distribution, both for

exhaustive and condensed summaries. In the second case, COPILS exploits appropriate distances and a matching step to output efficiently tractable comparative summaries. Future works will focus on a human-based qualitative evaluation of the interpretability of the generated differential summaries. As a criterion to measure the subjective notion of differential summary usefulness, a direction we consider relies on the capacity of a user to reconstruct the initial dataset from the summary of the modified one and the change description it provides.

References

1. Dong, G., Li, J.: Mining border descriptions of emerging patterns from dataset pairs. Knowledge and Information Systems **5**, 178–202 (2005)
2. Gemaque, R.N., Costa, A.F.J., Giusti, R., Dos Santos, E.M.: An overview of unsupervised drift detection methods. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery **10**(6), e1381 (2020)
3. Guillaume, S., Charnomordic, B., Loisel, P.: Fuzzy partitions: a way to integrate expert knowledge into distance calculations. Information sciences **245**, 76–95 (2013)
4. Kacprzyk, J., Zadrożny, S.: Protoforms of linguistic data summaries: Towards more general natural-language-based data mining tools. In: Abraham, A., del Solar, J.R., Koeppen, M. (eds.) Soft computing systems, pp. 417–425. Springer (2002)
5. Kacprzyk, J., Zadrożny, S.: Linguistic summarization of data sets using association rules. In: Proc. of the IEEE Int. Conf. on Fuzzy Systems. pp. 702–707. IEEE (2003)
6. Kacprzyk, J., Zadrożny, S.: Derivation of linguistic summaries is inherently difficult: Can association rule mining help? In: Borgelt, C., Gil, M.A., Sousa, J.M., Verleysen, M. (eds.) Towards Advanced Data Analysis by Combining Soft Computing and Statistics, pp. 291–303. Springer (2013)
7. Lesot, M.J., Smits, G., Pivert, O.: Adequacy of a user-defined vocabulary to the data structure. In: Proc. of the IEEE Int. Conf. on Fuzzy Systems. IEEE (2013)
8. Pilarski, D.: Linguistic summarization of databases with quantirius: a reduction algorithm for generated summaries. Int. Journal of Uncertainty, Fuzziness and Knowledge-Based Systems **18**(3), 305–331 (2010)
9. Smits, G., Yager, R.R., Pivert, O.: Interactive data exploration on top of linguistic summaries. In: Proc. of the IEEE Int. Conf. on Fuzzy Systems (2017)
10. Soulet, A., Crémilleux, B., Rioult, F.: Condensed representation of emerging patterns. In: Proc. of the 8th Asia Conf. on Knowledge Discovery and Data Mining (PAKDD04), LNCS, vol. 3056, pp. 127–132. Springer (2004)
11. Wilbik, A., Kacprzyk, J.: Towards an efficient generation of linguistic summaries of time series using a degree of focus. In: Proc. of the 28th North American Fuzzy Information Processing Society Annual Conf., NAFIPS'09 (2009)
12. Wilbik, A., Dijkman, R.M.: On the generation of useful linguistic summaries of sequences. In: Proc. of the IEEE Int. Conf. on Fuzzy Systems. pp. 555–562 (2016)
13. Wilbik, A., Kaymak, U., Dijkman, R.M.: A method for improving the generation of linguistic summaries. In: Proc. of the IEEE Int. Conf. on Fuzzy Systems (2017)
14. Wilbik, A., Keller, J.M.: A distance metric for a space of linguistic summaries. Fuzzy Sets and Systems **208**, 79–94 (2012)