Explaining Data Regularities and Anomalies

Amit K. Shukla[1], Grégory Smits[1], Olivier Pivert[1] and Marie-Jeanne Lesot[2]
[1] Univ Rennes, IRISA - UMR 6074, F-22305 Lannion, France
Email: {amit.shukla,gregory.smits,olivier.pivert}@irisa.fr
[2] Sorbonne Université, CNRS, LIP6, F-75005 Paris, France
Email: marie-jeanne.lesot@lip6.fr

***Abstract*—In the spirit of explainable AI approaches, this paper introduces a new strategy whose aim is to linguistically describe the inner structure of a dataset. Instead of removing irregular points and focusing on the analysis of regular points, the proposed approach relies on a unified data structure, an isolation forest, to both separate regular from irregular points and to identify their inner structure using a data-driven similarity measure. In addition, clusters of regular and irregular points are then linguistically described so as to help users focus on the most characteristic properties of each cluster and to possibly understand the reason why some points are irregular.**

I. Introduction

A first step toward the discovery of the knowledge embedded in a raw dataset is to exhibit its structure and to explain it to the user. In an unsupervised setting, clustering algorithms can be used to identify subgroups of points, called clusters, such that points assigned to the same cluster are more similar to one another than they are to points assigned to other clusters. This task relies on the definition of a comparison measure to determine how similar two points are. Points are grouped based on shared properties to form clusters. A cluster of a sufficient number of points describes a regular phenomenon that may be observed in the data. However, in many if not most of the applicative contexts, datasets cannot be completely and strictly decomposed into clusters of regular phenomena. There generally exist a few points that are very different from the others and that correspond to rare events, anomalies, outliers, etc. These points, that do not share enough data properties to be considered members of the identified clusters, are called irregularities in this work; they can affect the result of the clustering task if they are not dealt with through a specific process, as most clustering algorithms are not robust with respect to outliers. Moreover, being aware of the existence of these irregularities and understanding their properties is of a particular importance in a data-to-knowledge translation context.

In this work, we introduce a novel approach whose aim is to linguistically explain the structural properties of a dataset considering both regular and irregular points. Thanks to the output of the proposed strategy, the user precisely knows the main data properties of the found clusters of regularities, as well as the structural properties of the irregular points if they exist. Compared to existing approaches to outlier detection (see e.g [3]), our approach does not only assign a degree of irregularity to each suspicious point but also identifies properties shared by irregularities and links each cluster of anomalies with one (or several) cluster(s) of regularities.

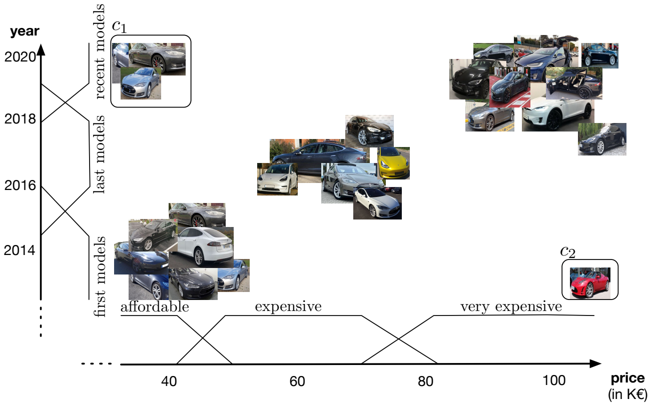

Fig. 1. Toy dataset of secondhand cars with regular trends and irregularities according to the price and year attributes.

To illustrate the goal and interest of the proposed approach, let us consider a user looking for a Tesla car to buy in a set of classified ads. Figure 1 depicts a fictitious distribution of the available cars according to their price and year. It may be observed that, when recent, these cars are very expensive and they may be considered expensive after a few years and only the oldest models are affordable. Wrt. these three groups of couples price/year that are generally observed, a few points do not follow these trends. The goal of this work is to explain the main trends that may be observed in the data as well as existing irregularities. It may for instance be interesting to underline the existence of the few recent Tesla (group c_1 on Fig. 1) that are cheap or the atypical car from 2014 (group c_2 on Fig. 1) that is still very expensive. Linguistic explanations are then provided to the user to give him/her a complete overview of the most interesting structural properties that may be observed in a dataset, concerning both regularities and irregularities.

To reach this goal, the approach proposed in this paper relies on an isolation forest [8], [5] and leverages the knowledge it captures for several subtasks: it first proposes to exploit it to compute a similarity measure that possesses the two interesting properties of being data-driven and contextual. It then proposes to use the isolation forest to identify irregularities and to exploit the derived similarity matrix as input to Agglomerative Hierarchical Clustering (AHC) to form clusters of regular and irregular points. It finally proposes to apply a

978-1-7281-6932-3/20/$31.00 ©2020 IEEE

linguistic rendering step to provide a linguistic description of the knowledge extracted from the data, both for the regular and irregular clusters, based on a fuzzy vocabulary also inferred from the data.

The rest of the paper is structured as follows. Section II first positions the proposed approach wrt. related existing works. Section III recalls the principle of an isolation forest, describes the proposed approach to infer a similarity matrix from it and how to use this similarity matrix within AHC, distinguishing between regular and irregular points. Section IV then details the linguistic rendering part of the approach. Experimentations described in Section V show the relevance and the efficiency of the proposed approach. Section VI concludes the paper.

II. Related Works

The functional goal of the proposed approach is to generate useful linguistic explanations about the data inner-structure. It thus belongs to the framework of linguistic summarization (see e.g. [1]), with the specificity to extract automatically the relevant and appropriate structure, based on the data cluster decomposition, beyond subsets of the data characterised by a fuzzy quantifier, as is usually the case with predefined protoforms. It can also be related to class characterization techniques that for instance focus on the linguistic explanation of the most typical points of each cluster [11]. Compared with this approach, the one proposed in this work differs on two main aspects.

The first one concerns the form of the generated explanations: a first possibility would be to consider all the possible conjunctive combinations of properties as possible explanations, which of course leads to more precise and discriminative descriptions of the classes but at the expense of a significant computation cost. Therefore, for an efficiency sake, it is proposed to quantify the extent to which each linguistic term from a fuzzy vocabulary is *typical* wrt. the considered class, considering the notion of typicality introduced in [9], as detailed in Section IV-B.

The second noteworthy aspect of the proposed approach is to consider that a dataset contains regular points but also irregularities that may also possess a cluster-based structure. In a context of fraud detection, the authors of [6] introduce an incremental strategy that makes it possible to exhibit such clusters of anomalies. In addition to being able to discover such clusters of rare and distant points, the approach proposed in this paper goes further, by identifying links between clusters of regular and irregular points and by incorporating a linguistic description of these links in the final summary.

This focus on the structural properties of the irregular points and their possible links with clusters of regular points also makes the proposed approach different from existing approaches dedicated to the identification of irregularities, as the Local Outlier Factor [2] or the isolation forest [8] and its extended version [5]. These classic approaches to anomaly detection compute anomaly scores to determine if points are regular or not. A quite opposite point of view is proposed in [14], where, instead of linguistically describing anomalies and their structure, these anomalies are detected from the linguistic description of the data using a specific distance measure. The granularity of the description provided by the approach proposed in this paper is finer since, in addition to being able to identify classes of anomalies, it also provides linguistic descriptions of the distinctive properties of each class.

III. Data Structuring based on an Isolation Forest

This section describes the approach proposed to determine the data structure, both in terms of regular and irregular clusters. As it relies on the exploitation of an Isolation Forest, the principle of this machine learning tool is first recalled. The following subsections describe the proposed similarity measure derived from an isolation forest and the proposed relational clustering procedure exploiting this similarity.

A. Reminder about Isolation Forest

In [8], an unsupervised technique to the identification of irregular points has been proposed, based on a principle called *isolation*. This approach leverages the property that irregular points are generally few and distant from regular points, whereas, on the contrary, regular points are numerous and form dense areas. The approach identifies anomalies using recursive isolation steps. An isolation step consists in randomly selecting an attribute and a threshold value between the minimal and maximal values taken by the points on this attribute. Points are then split into two subsets, depending on whether they take greater/lower values than the threshold for this attribute; the process is repeated on these subsets until all points have been isolated i.e. until each point is alone, thus in a leaf node. The output of the process is called an *isolation tree*. In order to ensure a sufficient coverage of the universe using random splits, a forest of such trees is built, called an *isolation forest*. Irregular points are those that appear in the top part of the isolation trees, as they correspond to points that can easily be isolated from the rest of the dataset.

A refinement of the concept of isolation forest has been proposed in [5] by considering affine functions (random selection of a slope and an intercept) instead of using separations that are parallel to the axes. This extension is used in the present work, as it leads to a better identification of the irregularities. Fig. 2 illustrates such an affine isolation tree for a 2D toy dataset.

Formally, given a set $\mathcal{D}$ of n data points, $\mathcal{D} = \{x_1, \dots, x_n\}$ from a universe U and described by p attributes A_1 to A_p, an isolation forest, denoted by $\mathcal{F}$, is composed of m isolation trees: $\mathcal{F} : \{T_1, \dots, T_m\}$, each tree being built on a randomly selected data subset.

An anomaly score, denoted by $aS(x)$, is computed for each point x based on the length of the paths from the root of each tree to x (the shorter these paths, the greater the anomaly score):

$$aS(x) = 2^{-\frac{E(P(x))}{\mathcal{N}}} \tag{1}$$

where $E(P(x)) \in (0, \max_{T \in \mathcal{F}} dep(T)]$ gives the average length of the paths from the root to x over all the constructed

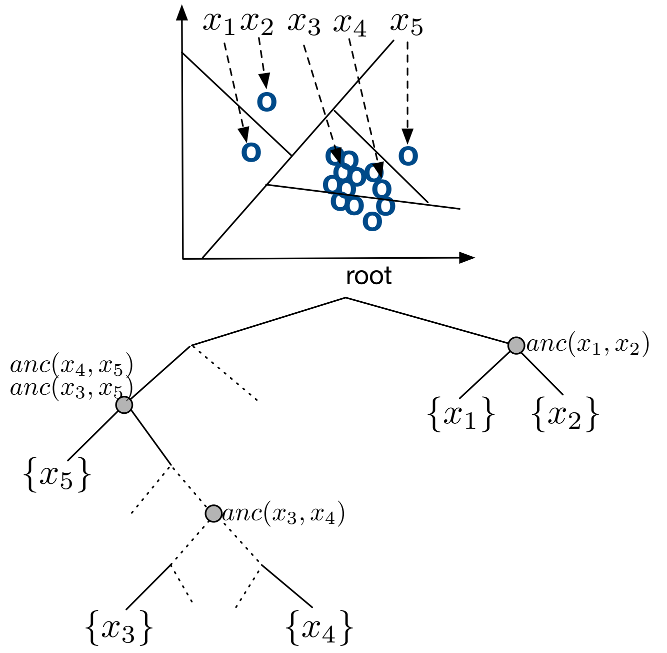

Fig. 2. (Top) 2D toy dataset with the considered affine separation functions and (bottom) excerpt of the associated induced isolation tree.

isolation trees, $dep(T)$ is the depth of tree T, and $\mathcal{N}$ is a normalization factor defined according to the size of the dataset (see [8] for more details).

B. Data-Driven Similarity using Isolation Forest

In the spirit of the similarity matrix inferred from a random forest introduced in [13], we propose to build, from an isolation forest, a similarity matrix that will then be used to cluster the data. Data clustering often relies on comparing the data using the Euclidean distance, although it may not respect the data characteristics or it may require to perform a preliminary attribute normalisation, e.g. needing expert knowledge about the relative attribute importance. Defining a data-driven measure intrinsically guarantees it is more appropriate to cluster the data. Moreover, as detailed below, the proposed measure offers the property of being contextual, which makes it rich and expressive.

The proposed measure leverages the knowledge captured by an isolation forest: the latter provides a unified model describing the data structure and embeds useful knowledge about the separability of the data points. Indeed, as mentioned in the previous subsection, points appearing in leaves on the top of the isolation trees correspond to points that are easy to isolate from others and that may be interpreted as irregularities. On the contrary, points that are isolated after a high number of separation steps are located in dense regions, pairs of points that remain unseparated after several splits are both close to each other and located in dense regions. We thus derive, as [13], the similarity between two points, say x and x', from the position in the trees of their deepest ancestor node that is denoted by $anc(x, x')$.

Formally, a similarity value $\text{sim}_T(x, x')$ is first computed for each tree T in the forest $\mathcal{F}$. The overall similarity between x and x', denoted by $\text{simIF}_{\mathcal{F}}(x, x')$, is then computed as the aggregation, by the average, of these values obtained over the whole isolation forest:

Definition 1. The similarity between two points x and x' according to an isolation tree T is computed as:

$$\text{sim}_T(x, x') = \begin{cases} 1 & \text{if } x = x' \\ \frac{\text{dep}(\text{anc}_T(x, x'))}{\text{dep}(T)} & \text{otherwise,} \end{cases} \tag{2}$$

where $\text{dep}(\text{anc}_T(x, x'))$ (resp. $\text{dep}(T)$) is the depth of the deepest common ancestor of x and x' (resp. the tree T).

Definition 2. The similarity between two points x and x' according to an isolation forest $\mathcal{F}$ is defined as:

$$\text{simIF}_{\mathcal{F}}(x, x') = \frac{1}{|\mathcal{F}|} \sum_{T \in \mathcal{F}} \text{sim}_T(x, x'). \tag{3}$$

where $|\mathcal{F}|$ is the total number of trees in the forest. The $\text{simIF}_{\mathcal{F}}$ measure has the classical properties of a similarity measure: it is normalized ($\text{simIF}_{\mathcal{F}}(x, x') \in [0, 1]$), symmetrical ($\text{simIF}_{\mathcal{F}}(x, x') = \text{simIF}_{\mathcal{F}}(x', x)$) and reflexive by design ($\text{simIF}_{\mathcal{F}}(x, x) = 1$).

Its main originality is to be *contextual*: two pairs of points at the same Euclidean distance, as for instance (x_1, x_2) and (x_3, x_4) in Figure 2, get different contextual similarities. Indeed, as the points x_3 and x_4 are located in a region of higher density than x_1 and x_2, they are split deeper in the isolation trees and get a higher similarity value. As a consequence, the clustering step may assign them to the same cluster, more than x_1 and x_2, which corresponds to a desired behavior. More generally, the contextual property of the proposed similarity measure is highly relevant in an unsupervised clustering setting as it avoids the separation of points that form dense areas.

C. Regularities and Irregularities Structuring using AHC

1) Clustering Step: Based on the isolation forest $\mathcal{F}$, the dataset $\mathcal{D}$ is first divided into two subsets $\mathcal{D}_I$ and $\mathcal{D}_R$ gathering irregular and regular points respectively. The separation criterion depends on their anomaly score $aS(x)$, as defined in Eq. 1 [8]:

$$\mathcal{D}_I = \{x \in \mathcal{D} / aS(x) > \gamma\},$$

and $\mathcal{D}_R = \mathcal{D} \backslash \mathcal{D}_I$. The γ parameter is a threshold whose value is discussed in Section V.

The similarity matrix is then computed using the $\text{simIF}_{\mathcal{F}}$ measure applied on $\mathcal{D}$. The datasets $\mathcal{D}_I$ and $\mathcal{D}_R$ are then clustered separately by means of Agglomerative Hierarchical Clustering (AHC), applied to the two similarity sub-matrices induced on each type of data. This strategy is motivated by the fact that, for the sake of interpretability, it appears preferable to process separately a large number of regular points and a few irregularities, since the latter would be more difficult to identify if all the data were processed together.

To build clusters of regular and irregular points from $\mathcal{D}_R$ and $\mathcal{D}_I$ respectively, the unweighted average linkage criterion

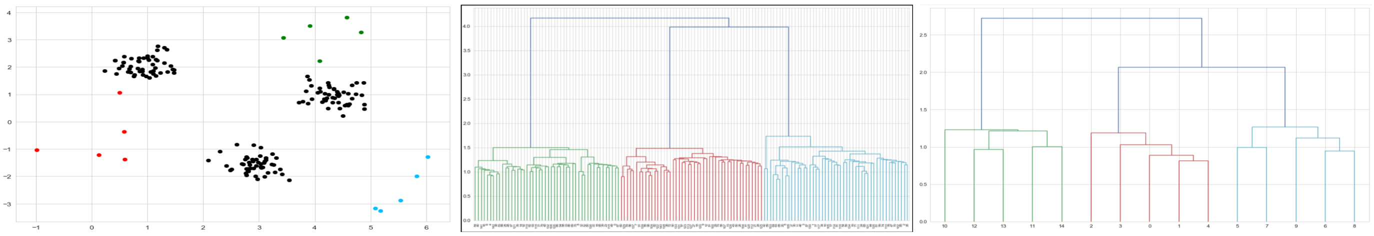

Fig. 3. (Left) Noisy dataset and clustering results, (middle) dendrogram of regular points, and (right) dendrogram on irregularities.

is used: the similarity between two groups of points is defined as the average of their pairwise similarities, formally $sim(C_1, C_2) = avg_{x \in C_1, y \in C_2} sim(x, y)$.

AHC outputs a hierarchy of data partitions, in the form of a dendrogram, that must be cut to derive the final cluster assignments. These dendrograms may be shown to the user to let him/her determine an appropriate partition of $\mathcal{D}_I$ and $\mathcal{D}_R$. In order to help the user in this task, we propose to use Dunn's index to automatically suggest one of the possible data partitions.

As a result of this clustering step, the datasets $\mathcal{D}_I$ and $\mathcal{D}_R$ are decomposed into k_I and k_R clusters, respectively denoted by $\mathcal{C}_I = \{c_I^1, \ldots, c_I^{k_I}\}$ and $\mathcal{C}_R = \{c_R^1, \ldots, c_R^{k_R}\}$.

2) Illustrative Example: Figure 3 illustrates the proposed two-step clustering process. The left part of the figure shows an artificial 2D dataset composed of three compact and dense clusters, that follow Gaussian distributions, and three groups of irregularities that were manually added. In a first step, irregular and regular points are separated based on the anomaly scores computed from the isolation forest. Then, the similarity matrix inferred from the isolation forest is used to build two dendrograms, one for the regular points (Fig. 3 middle) and one for the irregular ones (Fig. 3 right).

3) Relations between Outliers and Regular Clusters: In many applicative contexts, it is particularly important and relevant to identify links between clusters of irregular points and clusters of regular points. Irregularities may indeed correspond to anomalies when compared to expected or more frequent values, as for instance defects of a sensor. To identify such possible links, each cluster of irregularities is compared with each cluster of regular points. A similarity degree between a cluster c_I of irregularities and a cluster c_R of regularities, denoted by $simIF_{\mathcal{F}}(c_I, c_R)$, is computed as an average linkage:

$$simIF_{\mathcal{F}}(c_I, c_R) = \frac{1}{|c_I||c_R|} \sum_{\substack{x_I \in c_I \\ x_R \in c_R}} simIF_{\mathcal{F}}(x_I, x_R) \qquad (4)$$

It is worth recalling that the symmetrical similarity matrix is computed once and is stored in memory. The calculation of the above similarity degree between clusters thus simply consists in summing up some cells of a row in this matrix, with very little additional computational cost.

For each class of irregularities, it is thus possible to identify its closest class of regular points. Properties shared between the irregularities and this closest class of regular points, as well as their distinctive properties, are then identified and linguistically expressed, as detailed in the next section.

In this first part of the proposed approach, dedicated to the data clustering, only two hyperparameters have to be set: the number m of trees in the isolation forest and the anomaly threshold γ used to split the dataset into the two sets of regular and irregular points.

IV. Explanation of the Data Structure

Once the regular and irregular points have been structured into clusters, the next step of the proposed approach is to generate linguistic explanations exhibiting their characteristic properties, as well as the possible links between classes of regular and irregular points. To do so, for the sake of interpretability, numerical data properties are first translated into linguistic values using a fuzzy vocabulary, in a linguistic data rewriting step.

A. Linguistic Rewriting of the Data

1) Fuzzy Vocabulary: In order to describe the characteristic properties of each cluster in an interpretable way, a linguistic vocabulary is used. Formally, this vocabulary, denoted by $\mathcal{V} = \{V_1, \ldots, V_p\}$, consists of a set of linguistic variables, associated with each attribute: V_j is a triple $\langle A_j, \{v_{j1}, \ldots, v_{jq_j}\}, \{l_{j1}, \ldots, l_{jq_j}\}\rangle$ where q_j denotes the number of modalities associated with attribute A_j, the v_j's denote their respective membership functions defined on the domain of attribute A_j and l_{js} their respective linguistic labels, generally adjectives from the natural language. For instance, an attribute A describing prices may be associated with $q_A = 3$ modalities, in turn associated with the labels $l_{A1} =$ 'affordable', $l_{A2} =$ 'reasonable' and $l_{A3} =$ 'expensive'.

It is assumed that the linguistic variables associated with an attribute, say A_j, define a strong fuzzy partition [10]: $\forall y$, $\sum_{s=1}^{q_j} v_{js}(y) = 1$. As a consequence, any value y can be rewritten in terms of V and y can partially satisfy only up to two modalities, that, in addition, are adjacent.

Defining such a vocabulary is not an easy task for an end user, this is why an initial version of this vocabulary is automatically derived from the data, regarding the relevant fuzzy modalities needed for a good description of the (cluster-based) structure of the data. Of course, the end-user in charge of the data-to-knowledge inference process has to have a good understanding of these fuzzy partitions, and probably to provide the associated linguistic labels. The approach introduced in [12] is used to automatically infer an initial definition of the fuzzy vocabulary from the data, ensuring a good adequacy between the fuzzy partitions and the data structure. The user may then revise some parts of the vocabulary and assign his/her own linguistic labels to each modality.

2) Cluster Linguistic Description: For each cluster, say c, a fuzzy set of linguistic values taken from $\mathcal{V}$ is computed. This set, denoted by $\mathcal{L}_c$, provides knowledge about the coverage of the points assigned to c by each term of the vocabulary. It is called the *linguistic description* of c. It is formally defined as follows:

$$\mathcal{L}_c = \{v \mid \rho_v(c), v \in \mathcal{V}\},$$

The coverage of term v for cluster c, $\rho_v(c)$, is computed as the classical scalar cardinality (sigma-count) [4] of the fuzzy set v over cluster c: it is formally defined as

$$\rho_v(c) = \frac{1}{|c|} \sum_{x \in c} \mu_v(x). \tag{5}$$

B. Identification of Characteristic Properties

1) Typicality Degrees Computation: In order to identify the characteristic terms to describe each cluster, we propose to use the framework of typicality [9], [7], and to define characteristic properties as typical ones: a linguistic term is relevant to describe a cluster if it is both shared by the members of the cluster and if it does not apply to members of other clusters. The first constraint can be interpreted, in the typicality framework, as the notion of internal resemblance and the second one as external dissimilarity.

To measure the extent to which a term is shared by the members of the considered cluster, we propose to use the coverage measure $\rho_v(c)$ defined in Eq. (5). To measure the extent to which it does not apply to members of the other clusters, we propose to use the complementary to 1 of its coverage of other clusters, ie $\rho_v(c')$ for $c' \neq c$ in the same type of clusters, ie $c' \in C_R$ if $c \in C_R$ and $c' \in C_I$ if $c \in C_I$. It is then needed to aggregate these values over c': the $\min$ (resp. $\max$) function would be too extreme to perform this aggregation, as a term would then be considered specific of c if it does not cover at least one (resp. none) of the other clusters c'. A compromise looks preferable, therefore we propose to use an OWA operator to combine the $\rho_v(c')$'s, requiring that v does not cover most of the other clusters. More precisely, as detailed in Section V, higher weights are assigned to low values of $\rho_v(c')$'s so as to tip the scale in favor of a conjunctive behaviour.

Finally, these two quantities must be aggregated: we propose to apply a conjunctive aggregation operator, so as to express the double requirement that the term is shared by the members of the considered cluster and does not apply to the other clusters. The typicality degree of term v for cluster c is thus defined as

$$\tau_v(c) = \min(\rho_v(c), 1 - \text{OWA}_{c' \neq c}(\rho_v(c'))),$$

Considering a cluster c and its linguistic description $\mathcal{L}_c$ wrt. a vocabulary $\mathcal{V}$, a fuzzy set denoted by $\mathcal{T}_c$ is built to gather the typical properties of c; this fuzzy set contains the linguistic terms that best describe cluster c and is defined as:

$$\mathcal{T}_c = \{v \mid \tau_v(c), v \in \mathcal{V}\}.$$

2) Linguistically Explaining Irregular Clusters: Each cluster of irregularities c_I is associated with its closest cluster of regular points c_R, as described in Section III-C3, $c_R = \arg\max_{c \in \mathcal{C}_R} \text{simIF}_{\mathcal{L}}(c_I, c)$. If the corresponding similarity value is high enough, cluster c_I appears to gather anomalies of cluster c_R . The next step is then to identify the properties shared by c_I and c_R, as well as the ones that make c_I different from c_R (i.e., the reasons why the points from c_I are anomalies wrt. to c_R). The set of terms that are shared by c_I and c_R is denoted by $\mathcal{M}(c_I, c_R)$. It is computed by a set intersection operation:

$$\mathcal{M}(c_I, c_R) = \{v \mid \mu_{c_I \cap c_R}(v), v \in \mathcal{L}_{c_I} \cap \mathcal{L}_{c_R}\} \tag{6}$$

where $\mu_{c_I \cap c_R}(v) = \min(\rho_v(c_I), \rho_v(c_R))$.

The set of terms that make c_I a cluster of anomalies wrt c_R is denoted by $\mathcal{A}(c_I, c_R)$. It is computed by a set difference operation:

$$\mathcal{A}(c_I, c_R) = \{v \mid \mu_{c_I - c_R}(v), v \in \mathcal{L}_{c_I}\} \tag{7}$$

where $\mu_{c_I - c_R}(v) = \min(\rho_v(c_I), 1 - \rho_v(c_R))$.

Based on the set of characteristic terms associated with each cluster as well as the links identified between irregular and regular points, it is then possible to provide a detailed description of the noteworthy structural properties that may be observed in the considered dataset.

C. Linguistic Explanations

To make the data structure comprehensible to the end user, linguistic explanations are generated: they are illustrated in Section V describing experimental results and here defined formally. The first objective is to exhibit the typical terms of each cluster. So, for each cluster c of regular or irregular points, its set of typical terms is used to generate linguistic statements of the form:

Typical properties of c are:

- *A is v*
- ...

The properties are displayed from the most typical to the least typical one, i.e. in descending order of their $\tau_v(c)$ degree. A threshold may be used to filter out the terms that are not representative enough. Another option is to integrate these degree values if the user prefers more information rather than a purely linguistic summary: the former may be considered more legible, but less informative.

If a cluster of irregularities, say c_I, is interpreted by the proposed methodology as associated with the cluster of regular points c_R, the following linguistic explanations are added after the description of c_R:

The irregular points from c_I may be anomalies of c_R as:

- *they share the following characteristic properties:*
  - *A is v,*
  - ...
- *but they differ on:*
  - *A' is v'*
  - ...

where the properties shared (resp. not shared) by c_I and c_R are shown in decreasing order of their attached matching degree $\mu_{c_I \cap c_R}(v)$ (resp. difference degree $\mu_{c_I - c_R}(v)$). Again, the numerical values may be integrated in the summary or not, depending on the user preferences.

V. Experimental Study

This section gives the results of experimentations conducted on three sets of artificial data. They confirm the relevance of the similarity relation captured directly from the data and the informativity of the explanations we provide about the data structure.

A. Experimental Protocol

1) Datasets: Figure 4 shows the three datasets used to assess the proposed approach. Regular points are generated so as to follow Gaussian distributions and thus to form compact and dense clusters. Irregular points are then manually added in such a way that they may be structured by means of clusters and that some of these clusters are close enough to classes of regular points to be considered as their anomalies.

In the first dataset called $\mathcal{D}_1$ (left side of Figure 4) six classes are considered, three of regular points $\mathcal{C}_R = \{c_1, c_4, c_5\}$ and three classes of irregular points $\mathcal{C}_I = \{c_2, c_3, c_6\}$. $\mathcal{D}_2$ (Fig. 4 middle) has a similar structure: three clusters of high density can be considered as regularities $\mathcal{C}_R = \{c_2, c_4, c_6\}$, and three irregular clusters $\mathcal{C}_I = \{c_1, c_3, c_5\}$. Dataset $\mathcal{D}_3$ (Fig. 4 right) is, structurally speaking, more tricky to describe. It corresponds to a classic clustering data benchmark. A possible cluster decomposition of $\mathcal{D}_3$, admittedly subjective, is suggested on the figure, consisting in seven clusters, four describing dense regions $\mathcal{C}_R = \{c_2, c_3, c_5, c_7\}$ and three that may be interpreted as irregularities $\mathcal{C}_I = \{c_1, c_4, c_6\}$, c_1 (resp. c_4) being composed of points surrounding c_2 (resp. c_3). Sparse points surrounding c_5 with regular spacing are considered as a group of anomalies (c_6).

The cluster decomposition of three datasets as well as the proposed partitions are obviously debatable and subjective, but they are only used to compare a possible expected result and the one generated by the proposed approach.

For each dataset, the approach introduced in [12] is used to infer a fuzzy vocabulary that matches its class-based structure. These vocabularies are also depicted in Figure 4.

2) Hyperparameters: The first parameter is the number m of trees determining the size of the isolation forest. We follow the recommendation from [8], that is based on a thorough experimental study, and set $m = 100$. The second parameter is the anomaly detection threshold γ, that we set as $\gamma = 0.53$, i.e. slightly higher than the 0.5 value recommended in [8]. This values makes it possible to identify correctly the anomalies for all considered datasets.

Only one parameter is specific of the proposed approach, the weights used by the OWA operator to compute the typicality degree of each term appearing in the linguistic description of a cluster. Now, to be representative, a term has to sufficiently cover the concerned cluster and not so much the others. It is considered in this work that to conclude about the non representativity of a term wrt. a set of clusters, the term should cover only a minority of them. Thus, denoting s the number of considered clusters, we propose to set an equal weight equal to $\frac{2}{s}$ to all the $\frac{s}{2}$ clusters having the lowest $\rho_v(c')$ degrees.

	$\mathcal{D}_1$	$\mathcal{D}_2$	$\mathcal{D}_3$
nb. of irregularities	15	10	96
precision	1	0.43	0.88
recall	1	1	0.78

TABLE I

Accuracy of the irregularity detection step, using $\gamma = 0.53$.

	$\mathcal{D}_1$	$\mathcal{D}_2$	$\mathcal{D}_3$
ARI $\mathcal{D}_I$	1	1	0.55
ARI $\mathcal{D}_R$	1	1	0.99

TABLE II

ARI of the built partitions of regular and irregular clusters.

B. Results and Interpretations

1) Numerical Assessment: The first step of the approach is to separate irregularities from regular points using the anomaly score computed from the isolation forest. Even if the goal of these experimentations is not to show again the efficiency of an isolation forest-based approach to perform this task, Table I gives the result of this dataset decomposition into two subsets $\mathcal{D}_R$ and $\mathcal{D}_I$. The row *nb. of irregularities* corresponds to the number of points manually labelled as irregularities. For dataset $\mathcal{D}_1$ (resp. $\mathcal{D}_2$), it corresponds to the manually added irregular points forming classes c_2, c_3 and c_6 (resp. c_1, c_3 and c_5) in Fig. 4 left (resp. center). For $\mathcal{D}_3$, we consider that the points surrounding c_2, c_3 and c_5 are irregularities. Even if the irregularities around c_5, labelled c_6, may be considered a cluster of regular points, this cluster will then be reconstructed as a class of irregularities. Precision and recall of the irregularity detection step are given in Table I. They show satisfactory results except for $\mathcal{D}_2$ precision, which is due to the fact that the irregular cluster c_5 is not identified as such (see the visualisation discussed below).

Considering the labels given to the regular and irregular points, Table II assesses the quality, measured by the Adjusted Rand Index, ARI, of the partition produced by AHC with cut parameter set so as to have the number of clusters that optimises the Dunn's index. It shows that the obtained clusters match the expected ones, except for the irregular points in dataset $\mathcal{D}_3$. This can be explained by the chosen ground-truth to assign all points in c_6 to a single cluster, as previously discussed.

2) Visualisation: The results obtained for $\mathcal{D}_1$ are shown in Figure 3 and commented in Section III-C. Regarding $\mathcal{D}_2$, the results are depicted in Figure 5. The first two figures show the two subsets $\mathcal{D}_R$ and $\mathcal{D}_I$. It may be observed that some points labelled as regularities are considered irregularities due to their peripheral position wrt. the dense region of the clusters of regular points, hence the perfectible precision given in Table I. The two figures on the right part of Figure 5 give the dendrograms of these two subsets $\mathcal{D}_R$ and $\mathcal{D}_I$.

Concerning $\mathcal{D}_3$, Figure 6 also gives the separation between regular (left) and irregular (center left) points obtained using

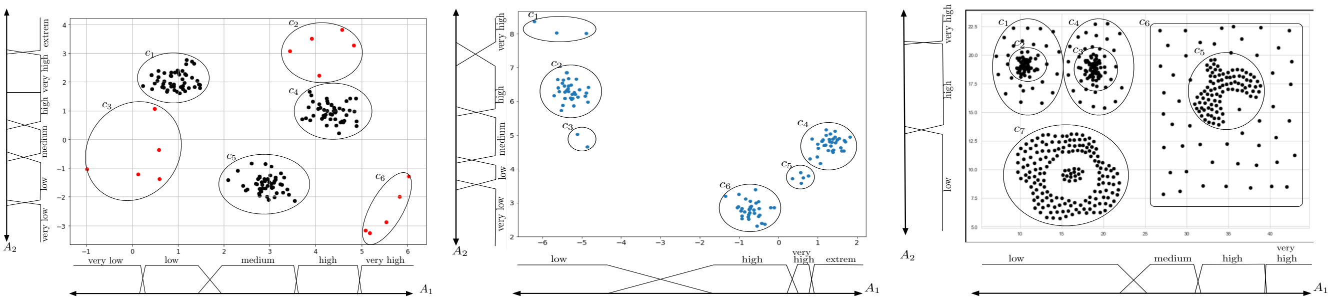

Fig. 4. Artificial datasets and their automatically generated fuzzy vocabularies: $\mathcal{D}_1$ (left), $\mathcal{D}_2$ (center) and $\mathcal{D}_3$ (right).

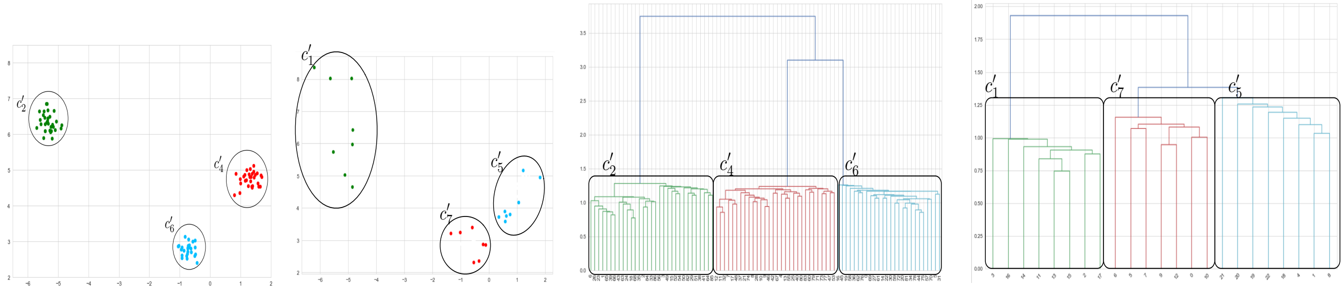

Fig. 5. Decomposition of $\mathcal{D}_2$ in $\mathcal{D}_R$ (left) and $\mathcal{D}_I$ (center left), and their associated dendrograms (center right for $\mathcal{D}_R$ and right for $\mathcal{D}_I$).

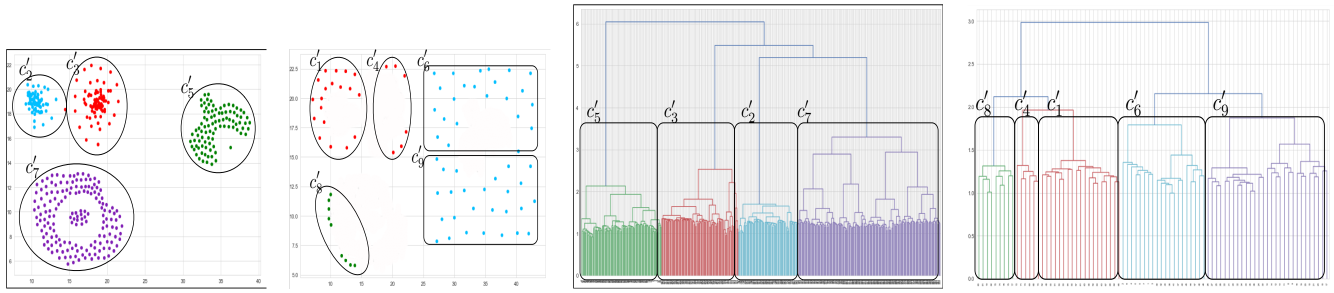

Fig. 6. Decomposition of $\mathcal{D}_3$ in $\mathcal{D}_R$ (left) and $\mathcal{D}_I$ (center left), and their associated dendrograms (center right for $\mathcal{D}_R$ and right for $\mathcal{D}_I$).

the anomaly degree (Eq. 1). Compared with the subjective decomposition of $\mathcal{D}_3$ proposed in Figure 4, irregularities found by the isolation forest are a little bit different. Cluster c_6 is correctly identified but some points between c_2 and c_3 are considered regularities. Some points in the peripheral of c_7 are also considered irregularities. Whereas the dendrogram computed on $\mathcal{D}_R$, shown on Figure 6, correctly identifies the different clusters of regular points, which confirms the relevance of the similarity matrix inferred from the isolation forest, the dendrogram built on $\mathcal{D}_I$ needs a deeper study to be understood. The green points that correspond to the irregularities surrounding c_7 are easily grouped, as well as the points surrounding the clusters c_2 and c_3 that are well split into two groups then gathered in the top of the dendrogram (cut at 1.9). However, the points that correspond to cluster c_6 in Figure 4 are divided into two groups (blue and purple) that are gathered on the very top of the dendrogram (cut 2.2). This is due to the use of the average linkage criterion that separates the points located on top of c_5 from those under c_5.

3) Explaining Irregular Clusters: To ease the understanding of the class-based structure of each dataset, a similarity degree is then computed between the clusters identified in $\mathcal{D}_I$ and those from $\mathcal{D}_R$. The results of these comparisons for the clusters labelled in Figures 5 and 6 are given in Table III. The similarities found between these classes provide useful additional information about the partitions provided by AHC as it makes it possible to better understand the reasons why points are considered irregularities and to identify false positive. For instance, concerning $\mathcal{D}_3$, c'_2 is linked with c'_1, c'_3 with c'_4, c'_7 with c'_8, and finally c'_5 is linked with c'_6 and c'_9.

4) Linguistic Summaries: The final step of the approach is to provide the user with explanations about the structural properties of the classes found in each subset $\mathcal{D}_R$ and $\mathcal{D}_I$, as well as explanations about the links that exist between irregularities and classes of regular points. To show the kind of linguistic explanations provided, below are given the explanations about $\mathcal{D}_3$ and more precisely about the class c'_5 and its irregularities grouped in c'_6 and c'_9:

$\mathcal{D}_1$	c_1	c_4	c_5
c_2	0.0195	**0.0401**	0.0129
c_3	**0.0676**	0.0206	0.0228
c_6	0.0101	0.0206	**0.0317**

$\mathcal{D}_2$	c'_2	c'_4	c'_6
c'_1	**0.0377**	0.0116	0.0209
c'_7	0.0093	0.0314	**0.0788**
c'_5	0.0086	0.0484	**0.0586**

$\mathcal{D}_3$	c'_2	c'_3	c'_5	c'_7
c'_1	**0.01865**	0.01567	0.00552	0.01403
c'_4	0.00881	**0.02447**	0.0107	0.01159
c'_6	0.00392	0.00808	**0.01657**	0.00433
c'_8	0.00535	0.00732	0.00564	**0.02556**
c'_9	0.0023	0.00476	**0.01539**	0.00444

TABLE III
SIMILARITY MATRICES BUILT BETWEEN CLASSES OF IRREGULARITIES AND REGULARITIES USING THE $simIF_{\mathcal{F}}(c_i, c_r)$ MEASURE (EQ. 4).

Typical properties of c'_5 in $\mathcal{D}_3$ are:

- *A_1 is high*
- *A_2 is high*
- *The irregular points from c'_6 may be anomalies of c'_5 as:*
  - *they share the following characteristic properties:*
    - *A_1 is high*
    - *A_2 is high*
  - *but they differ on:*
    - *A_1 is medium*
    - *A_1 is very high*
- *The irregular points from c'_9 may be anomalies of c'_5 as:*
  - *they share the following characteristic properties:*
    - *A_1 is high*
  - *but they differ on:*
    - *A_2 is low*
    - *A_1 is medium*
    - *A_1 is very high*

This linguistic description provides a legible summary of the data structure, both in terms of regular and irregular points.

VI. CONCLUSION AND PERSPECTIVES

Faced with a new dataset to analyze, users need tools to ease the translation of data into knowledge. Before deciding to conduct a deep and costly analysis of these data, it may be of particular interest to have explanations, automatically generated, about the noteworthy structural properties that may be observed in the data. This work introduces a novel and pragmatic approach whose aim is to first automatically identify classes of regular and irregular points as well as links between them, for instance to explain the provenance of irregularities that thus could be interpreted as anomalies. A particularity of this clustering process is to rely on a unified data structure to both quantify the similarity between points as well as to differentiate between regular and irregular points: an isolation forest. The distinctive properties of each class as well as those that make of a set of irregular points possible anomalies of a set of regular points are then linguistically explained to the user. We have shown on three datasets the relevance of the knowledge automatically acquired from the data.

The promising results obtained with this new strategy to knowledge extraction from data that focus on the comparison between observable regular trends and irregularities open several perspectives of future works. We are conducting further experimentations to compare the proposed approach with other approaches dedicated to anomaly detection in a fraud detection challenge (IEEE-CIS Fraud detection challenge https://www.kaggle.com/c/ieee-fraud-detection). To make the process faster, we are also working on a distributed computation of the isolation forest and the similarity matrix that can be inferred from it.

ACKNOWLEDGEMENTS

This research takes part of the CREDOC project partially founded by the Mer Bretagne Atlantique cluster and the Image and Network cluster.

REFERENCES

[1] F. E. Boran, D. Akay, and R. R. Yager. An overview of methods for linguistic summarization with fuzzy sets. *Expert Systems with Applications*, 61:356–377, 2016.

[2] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander. Lof: identifying density-based local outliers. In *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, pages 93–104, 2000.

[3] V. Chandola, A. Banerjee, and K. Kumar. Anomaly detection: A survey. *ACM Computing Surveys*, 41(3):1–58, 2009.

[4] D. Dubois and H. Prade. Fuzzy cardinality and the modeling of imprecise quantification. *Fuzzy sets and Systems*, 16(3):199–230, 1985.

[5] S. Hariri, M. C. Kind, and R. J. Brunner. Extended isolation forest. *arXiv preprint arXiv:1811.02141*, 2018.

[6] M.-J. Lesot and A. Revault d'Allonnes. Credit-card fraud profiling using a hybrid incremental clustering methodology. In *Proc. of the Int. Conf. on Scalable Uncertainty Management (SUM'12)*, 2012.

[7] M.-J. Lesot, M. Rifqi, and B. Bouchon-Meunier. Fuzzy prototypes: From a cognitive view to a machine learning principle. In *Fuzzy Sets and Their Extensions: Representation, Aggregation and Models*, pages 431–452. Springer, 2008.

[8] F. T. Liu, K. M. Ting, and Z.-H. Zhou. Isolation-based anomaly detection. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 6(1):1–39, 2012.

[9] M. Rifqi. Constructing prototypes from large databases. In *Proc. of the Int. Conf. on Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU'96)*, pages 300–306, 1996.

[10] E. H. Ruspini. A new approach to clustering. *Information and Control*, 15(1):22 – 32, 1969.

[11] G. Smits and O. Pivert. Linguistic and graphical explanation of a cluster-based data structure. In *Scalable Uncertainty Management*, pages 186–200. Springer, 2015.

[12] G. Smits, O. Pivert, and M.-J. Lesot. Vocabulary elicitation for informative descriptions of classes. In *2017 Joint 17th World Congress of International Fuzzy Systems Association and 9th International Conference on Soft Computing and Intelligent Systems (IFSA-SCIS)*, pages 1–8. IEEE, 2017.

[13] K. Torkkola and E. Tuv. Ensemble learning with supervised kernels. In *European Conference on Machine Learning*, pages 400–411. Springer, 2005.

[14] A. Wilbik and J. M. Keller. Anomaly detection from linguistic summaries. In *2013 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pages 1–7. IEEE, 2013.