On Dissimilarity Measures at the Fuzzy Partition Level

Grégory Smits, Olivier Pivert, Toan Ngoc Duong

On Dissimilarity Measures at the Fuzzy Partition Level

Grégory Smits, Olivier Pivert and Toan Ngoc Duong

IRISA - University of Rennes, UMR 6074, Lannion, France
{gregory.smits | olivier.pivert | ngoc-toan.duong}@irisa.fr

Abstract. On the one hand, a user vocabulary is often used by soft-computing-based approaches to generate a linguistic and subjective description of numerical and categorical data. On the other hand, knowledge extraction strategies (as e.g. association rules discovery or clustering) may be applied to help the user understand the inner structure of the data. To apply knowledge extraction techniques on subjective and linguistic rewritings of the data, one first has to address the question of defining a dedicated distance metric. Many knowledge extraction techniques indeed rely on the use of a distance metric, whose properties have a strong impact on the relevance of the extracted knowledge. In this paper, we propose a measure that computes the dissimilarity between two items rewritten according to a user vocabulary.

Keywords: Fuzzy partition, data personalization, dissimilarity measure

1 Introduction

Helping users extract and understand the content of a raw data set is a crucial task in data mining. Most of the datasets contain the description of items on attributes that are generally of a numerical or a categorical nature. It is cognitively difficult for an end-user to browse and analyze a large collection of numerical and categorical data, and it is moreover, technically speaking, almost impossible to generate an interpretable graphical view of a set of data described on more than 3 dimensions. To overcome these difficulties, soft-computing-based approaches of data management leverage a user vocabulary to turn numerical and categorical variables (i.e. attributes) into linguistic variables. Once rewritten according to the user vocabulary, concise and easily interpretable views of the data may be generated to give the user an insight into the content of the dataset [1]. In addition, data mining techniques, as clustering algorithms for instance, may be used to discover the inner structure of the data, whose description also constitutes valuable knowledge [2]. Many data mining techniques rely on a distance measure to determine the similarity of two items. In this work, we address the question of computing the distance between two items rewritten according to a user vocabulary formalized by means of strong fuzzy partitions. This question of a distance measure at the partition level has been notably studied by Guillaume et al. [3], but the measure they proposed sometimes leads to questionable results

as we will see in Section 2.4.

We propose in this paper a new dissimilarity measure at the partition level that somehow reconsiders the indistinguishability relation introduced by the use of a fuzzy vocabulary for the sake of a better interpretability of the generated results. The final objective is to use the proposed dissimilarity measure to build clusters of data rewritten according to a user vocabulary instead of considering their numerical and categorical values. Motivation for that are manifold. First, the indistinguishability area defined by the cores of the fuzzy sets will reduce the number of distinct rewritings to consider, thus making it possible to handle larger datasets. Second, translating numerical and categorical values into linguistic terms allows for the conception of graphical views representing the obtained clusters on many dimensions at the same time [1], which cannot be envisaged on numerical/categorical data. And third, starting with a rewriting step of the data is a way to personalize the data-to-knowledge translation process and to make it more easily interpretable for end-users. But the relevance of the structure built by a clustering highly depends on the properties of its underlying distance measure.

In this paper, we focus on the definition of dissimilarity measures at the fuzzy partition level and the study of their properties. Their use by a clustering process will be the next step. The rest of the document is structured as follows. In Section 2, preliminary notions regarding fuzzy-set-based vocabularies and dissimilarity measures at the partition level are recalled. Sections 3.1 and 3.2 detail our proposed dissimilarity measures, respectively for numerical and categorical domains.

Motivating Example

To illustrate the motivation for a new dissimilarity measure, let us consider the vocabulary, i.e. fuzzy partition, illustrated in Fig. 1 that turns the mileage of a car into a linguistic variable that may take the values {veryLow, low, medium, high, veryHigh}. In the situation illustrated by Fig. 1, a dissimilarity measure at the partition level has to be able to capture the fact that t_1 is closer to t_3 than to t_6 because the linguistic value that describes t_1, namely *low mileage*, is closer to *medium mileage* than to *veryHigh mileage*, this case being well covered by the measure defined in [3]. However, contrary to [3], we argue that the indistinguishability relation should be limited to the core of the fuzzy sets (as e.g. between t_3 and t_4), and that it appears more natural and interpretable to consider t_3 as closer to t_2 than to t_5 even if these last two points satisfy the linguistic value *medium mileage* at the same degree (in this case $\mu_{medium}(t_2) = \mu_{medium}(t_5) = 0.7$). This expected behavior is all the more important if the considered task is to build groups of items having close linguistic rewritings. Using a dissimilarity measure that is more appropriate to compare rewritten data, we expect that more meaningful groups of items will be obtained especially by avoiding grouping tuples that are significantly different.

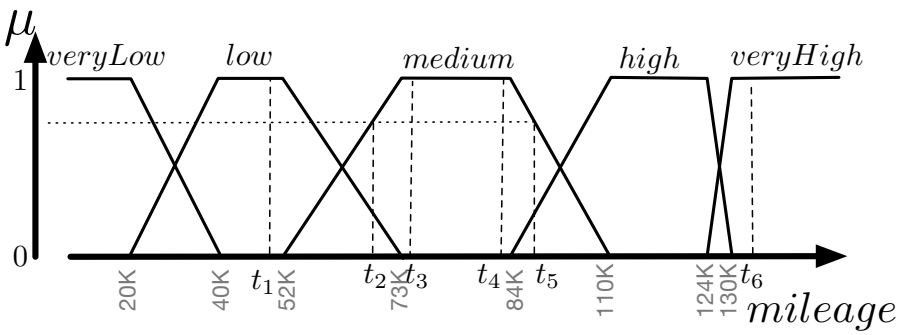

Fig. 1. Distance computation at the partition level between numerical values

Even if meaningful fuzzy partitions may be built on categorical attributes, using a dedicated graphical interface as ReqFlex for instance [4] (Fig. 2), distance measures between categories or discrete fuzzy sets are generally reduced to a Boolean test of equality. A second contribution of this paper is to propose a measure to compute the dissimilarity between two categorical values that takes into account the structure of its underlying user vocabulary. The idea is to consider that discrete fuzzy sets sharing some categories should be considered as somewhat semantically related. By doing so, one may infer a weak partial order on discrete fuzzy sets defined on top of a categorical attribute. Thus, categorical values taken from these two sets should be considered closer to each other than categorical values taken from two sets having an empty intersection. To illustrate this proposal, let us consider a possible fuzzy vocabulary describing different car brands according to their relative reliability reputation (Fig. 2). Then, we argue that the brand *Chrysler*, characterized as a fully *moderatelyReliable brand*, should be considered as closer to *VW*, a *reliable brand*, than *Daewoo*, that belongs to the set of *poorlyReliable brands*, because *moderatelyReliable brands* and *reliable brands* have in this case much more in common than with *poorlyReliable brands*. Obviously, the relevance of this interpretation of semantic closeness is context-dependent, and most of all depends on the point-of-view expressed by the user through the definition of his/her vocabulary.

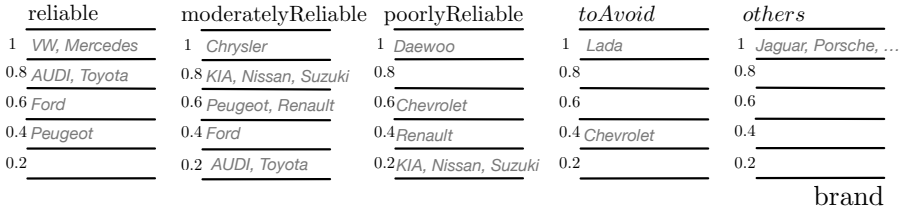

Fig. 2. Example of a subjective vocabulary on a categorical attribute

2 Preliminary notions

Let $\mathcal{D} : \{t_1, t_2, \ldots, t_m\}$ be a set of m items to analyze. Each item is initially defined by the values it takes on n attributes $\{A_1, A_2, \ldots, A_n\}$ that may be of a numerical or categorical type. More formally, if one denotes by X_i the definition domain of attribute A_i then $t \in X_1 \times X_2 \times \ldots \times X_n$. One denotes by $t.A$ the value taken by item t on attribute A.

2.1 Fuzzy-Set-Based User Vocabulary

We consider that a vocabulary composed of Fuzzy Partitions (FP) is defined on the attributes $\{A_1, A_2, \ldots, A_n\}$. Such a vocabulary, denoted by $\mathcal{V} = \{V_1, \ldots, V_n\}$, formally consists of a set of linguistic variables, associated with each attribute: V_j is a triple $\langle A_i, \{v_{i,1}, \ldots, v_{i,q_i}\}, \{l_{i,1}, \ldots, l_{i,q_i}\}\rangle$ where q_i denotes the number of modalities associated with attribute A_i, the v_i's denote their respective membership functions defined on domain X_i and the l_i's their respective linguistic labels, generally adjectives of the natural language. For instance, an attribute A_i describing prices may be associated with $q_i = 3$ modalities, in turn associated with the labels $l_{i,1}$ ='cheap', $l_{i,2}$ ='reasonable' and $l_{i,3}$ ='expensive'.

It is assumed that for all attributes, each value may be completely rewritten in terms of $V : \forall y \in D_j, \sum_{s=1}^{q_j} v_{js}(y) = 1$. Moreover, it is assumed that the partitions defined on numerical attributes form a strong FP [5], which leads to the constraint that y can partially satisfy up to two adjacent modalities. Figures 1 and 2 are examples of such partitions defined on a numerical and a categorical attribute respectively.

2.2 Item Rewriting vector

Initially defined in a numerical and categorical space, an item may be rewritten using the linguistic terms from the user vocabulary. The result of such a rewriting step is called an item rewriting vector.

Definition 1. *One denotes by R_t the rewriting vector of an item t wrt. a user vocabulary $\mathcal{V}$, this vector being the concatenation of the satisfaction degrees obtained by t on the different terms that compose $\mathcal{V}$. Such a vector is represented in the following way:*

$$R_t = \langle \mu_{v_{1,1}}(t), \mu_{v_{1,2}}(t), \ldots, \mu_{v_{1,q_1}}(t), \ldots, \mu_{v_{n,1}}(t), \mu_{v_{1,n}}(t), \ldots, \mu_{v_{1,q_n}}(t)\rangle.$$

We also denote by $R_t^{A_i}$ the part of the whole rewriting vector R_t that concerns the attribute A_i, $R_t^{A_i} = \langle \mu_{v_{i,1}}(t.A_i), \mu_{v_{i,2}}(t.A_i), \ldots, \mu_{v_{i,q_i}}(t.A_i)\rangle$.

Example 1. Tab. 1 shows the data (attribute values and rewriting vectors from Fig 1) that have to be considered when computing a dissimilarity at the FP level.

Table 1. Items from Fig. 1 and their rewriting vector

t	$t.mileage$	$R_t^{mileage}$	t	$t.mileage$	$R_t^{mileage}$
t_1	$50K$	$\langle 0, 1, 0, 0, 0\rangle$	t_2	$70K$	$\langle 0, 0.3, 0.7, 0, 0\rangle$
t_3	$74K$	$\langle 0, 0, 1, 0, 0\rangle$	t_4	$80K$	$\langle 0, 0, 1, 0, 0\rangle$
t_5	$90K$	$\langle 0, 0, 0.7, 0.3, 0\rangle$	t_6	$134K$	$\langle 0, 0, 0, 0, 1\rangle$

2.3 Properties of a Dissimilarity Measure at the Partition Level

When it comes to defining a dissimilarity that takes into account fuzzy sets, then three types of comparison may be envisaged [6]: 1) between two points that belong to a same fuzzy set, 2) between a point and a fuzzy set and, 3) between two fuzzy sets [7]. As shown in [3], (that is, to the best of our knowledge, the only existing approach addressing the question of a distance calculation at the fuzzy partition level) the measure we have to define has, in some sense, to combine these three types of fuzzy distances.

In fine, we aim at computing the dissimilarity between two items wrt. the considered vocabulary $\mathcal{V}$. This measure obviously relies on the aggregation of dissimilarities computed on each considered dimension. On a given dimension A_i, the dissimilarity at the partition level of two items, say t and t', has to combine the dissimilarity between the two numerical/categorical values ($t.A_i$ and $t'.A_i$) and between their rewriting wrt. $\mathcal{V}$: $R_t^{A_i}$ and $R_{t'}^{A_i}$. The expected behavior of the function to build is that the farther $t.A_i$ and $t'.A_i$, the higher the returned dissimilarity value. But, this function also has to take into account the indistinguishability relation embedded in the definition of a fuzzy subset, which means that the dissimilarity between $R_t^{A_i}$ and $R_{t'}^{A_i}$ should be 0 if $t.A_i$ and $t'.A_i$ fall in the core of a same partition element.

On any dimension involved in a rewriting vector, the function to define has to fulfil the following properties to constitute a *dissimilarity*:

- positiveness: $d(t, t') \geq 0$,
- identity of indiscernibles: a property that is generally defined in the following way $d(t, t') = 0 \Leftrightarrow t = t'$ but extended as follows in our particular context $d(t, t') = 0 \Leftrightarrow R_t = R'_t$ to capture the indistinguishability relation embedded in the FP,
- symmetry: $d(t, t') = d(t', t)$.

A dissimilarity that also satisfies the triangle inequality: $d(t, t') \leq d(t, t'') + d(t'', t')$, is called a semi-distance.

2.4 Behavior of Existing Approaches

In this subsection, we show that the existing approaches (a dedicated one [3] and a naive one) to the computation of a dissimilarity degree at the FP level lead, in some particular cases, to results difficult to understand and interpret.

A Generic Dissimilarity Measure Whatever the type of the attribute A_i concerned, numerical or categorical, a way to compute the dissimilarity of two items t and t', or more precisely their rewriting vectors $R_t^{A_i}$ and $R_{t'}^{A_i}$, is to simply compare one-by-one the respective membership degrees of t and t' on the different terms of the vocabulary. Such a dissimilarity measure, denoted by $d_1^i(t,t')$ may be formalized as follows:

$$d_1^i(t,t') = \frac{1}{q_i}\sum_{j=1}^{q_i} |\mu_{v_{i,j}}(t) - \mu_{v_{i,j}}(t')|.$$

The main advantage of this basic strategy is that it can be applied to both numerical and categorical attributes. However it suffers from the fact that it does not take into account the structure of the concerned FP. It indeed considers at the same distance of 1 any pair of values falling in the core of two distinct partition elements, whatever the position of these elements in the partition. In the example illustrated in Fig. 1, $d_1^{mile.}(t_1,t_3) = d_1^{mile.}(t_3,t_6) = d_1^{mile.}(t_1,t_6) = 1$.

A Pseudo-Metric at the FP-Level In [3], the authors address the question of distance calculation at the FP level, but for numerical attributes only. They especially define a pseudo-metric for the case of strong FP. This metric relies on a strict discretization of the universe of the concerned attribute as shown in Fig. 3 that form crisp areas denoted $\{I_1, I_2, \ldots, I_{q_i}\}$. Then, to compute the distance between two points on a given attribute A_i, their position within this discretization is first computed using the following function:

$$P(t) = I(t) - \mu_{v_i,I(t)}(t),$$

where $I(t)$ is the index of the area ($I(t) \in \{I_1, I_2, \ldots, I_{q_i}\}$) in which t is located.

Then, the dissimilarity is quantified by the function $d_2^i(t,t')$:

$$d_2^i(t,t') = \frac{|P(t) - P(t')|}{q_i - 1}.$$

Example 2. To illustre how dissimilarity degrees are computed using the measure d_2, let us consider the points t_2, t_4 and t_5 from Fig. 1. Then, these points are assigned to the following areas: $I(t_2) = I(t_4) = 3$ and $I(t_5) = 4$. Considering that $\mu_{v_{medium}}(t_2) = 0.7$ and $\mu_{v_{high}}(t_5) = 0.3$, we thus obtain the following distance degrees:

- $d_2^{mile.}(t_2,t_4) = \frac{|2.3-2|}{4} = 0.075$,
- $d_2^{mile.}(t_4,t_5) = \frac{|2-3.7|}{4} = 0.425$,
- $d_2^{mile.}(t_2,t_5) = \frac{|2.3-3.7|}{4} = 0.35$.

The metric d_2 handles well the distance between the partition elements to which the two points belong. If one goes back to the situation illustrated in Fig. 1, then $d_2(t_1,t_3) < d_2(t_1,t_6)$. However, despite the fact that the core of a

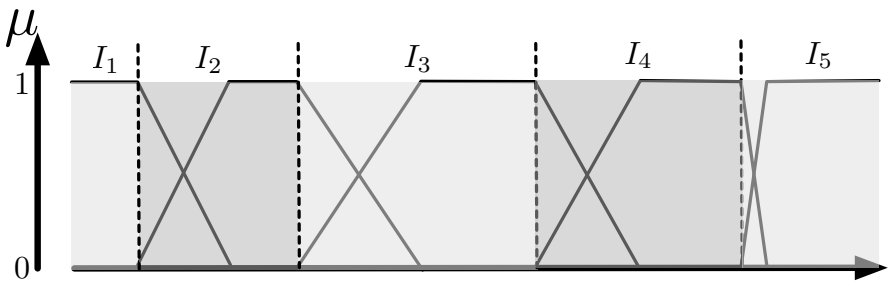

Fig. 3. Discretization of a numerical domain used by the metric d_2

partition element introduces an area of indistinguishability, it appears desirable to take into account the position of the points within the indistinguishability area when computing a distance with points outside this area. For the sake of understandability and interpretability, but also to improve the relevance of the data mining task that relies on a distance calculation, it indeed appears relevant and desirable to consider t_4 (Fig. 1) closer to t_5 than to t_2. However, in this particular case, as t_2 and t_4 fall in the same area according to the crisp discretization suggested in [3] ($I(t_2) = I(t_4) = 3$), then $d_2(t_2, t_4) < d_2(t_4, t_5)$, which, we think, is highly questionable.

3 A Dissimilarity at the FP Level

In this section, we propose a measure to compute the dissimilarity between two points that combines their respective position at the partition level as well as their value dissimilarity. In this sense, the proposed measure is inspired from works done in the context of distance calculation in image processing, and especially the fuzzy geodesic distance suggested in [8].

Definition 2. *We denote by $d_*(t, t')$ the global dissimilarity to determine between t and t' taking into account the structure of the FPs that form the vocabulary $\mathcal{V}$. $d_*(t, t')$ relies on the aggregation of dissimilarity degrees on the different considered dimensions, we thus denote by $d^i_*(t, t')$ the dissimilarity between t and t' on attribute A_i:*

$$d_*(t, t') = \frac{1}{n} \sum_{i=1}^{n} d^i_*(t, t'). \tag{1}$$

The functions d^i_*'s are defined in such a way that they return a dissimilarity degree in the unit interval, hence the co-domain of $d_*(t, t')$ is also $[0, 1]$. In the rest of this section, we provide definitions of $d^i_*(t, t')$, first when the concerned attribute is of a numerical type associated with strong FPs (Sec. 3.1), then when

it concerns a categorical attribute associated with a discrete fuzzy partition (Sec. 3.2).

3.1 For Numerical Attributes

We first address the question of computing the distance at the FP level between two points t and t' when the concerned attribute is of a numerical nature. To compute the distance between two values wrt. a strong FP, we consider the path formed by the boundaries of the partition elements that are above the line $y = 0.5$. As illustrated in Fig. 4, this path corresponds to the union of the convex hulls of each partition element. We denote by $\mathcal{L}_i$ this path for the partition V_i and $|\mathcal{L}_i|$ its length. A first strategy to define the limits of this path is to consider the minimum and maximum values present in the data on the concerned attribute. This strategy being very sensitive to extremum values, we propose a second one leveraging the fact that all the values inside the core of a partition element are indistinguishable. We thus consider that all the values fully satisfying the first (resp. last) element of the partition are at the same distance wrt. a point taken outside the core of this element. This allows us to consider that the path $\mathcal{L}_i$ starts with the right bound of the core of the first partition element and ends with the left bound of the core of the last element (See. Fig. 4). So every value inside the core of the first (resp. last) element of the partition is treated as the right (resp. left) bound of the core of the element in the dissimilarity calculation.

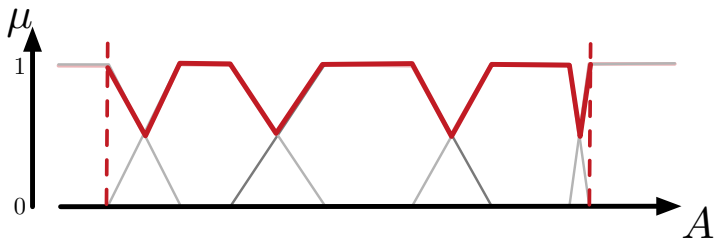

Fig. 4. Hull of a strong FP

To compute the dissimilarity between two points wrt. a strong FP, we then distinguish between two cases. When the two values to compare fall in the core of a same modality, then we assume their distance to be equal to 0 so as to satisfy the indistinguishability relation introduced by the different fuzzy sets. In all other cases, the distance between two values corresponds to the length of the path following $\mathcal{L}_i$ between these two values. Such a path between two values, say t and t'[1], is denoted by $\mathcal{L}_i(t, t')$ as illustrated in Fig. 5.

[1] For the sake of simplicity, t and t' are used instead of $t.A_i$ and $t'.A_i$ respectively to lighten the notation.

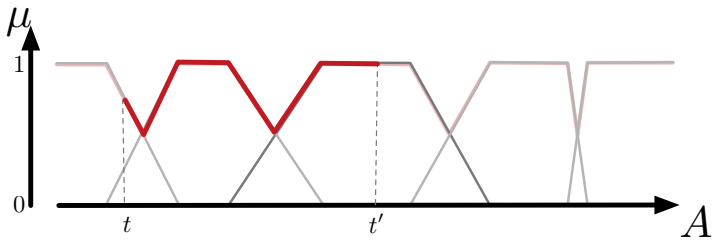

Fig. 5. Path between two values t and t'

Definition 3. *Let A_i be a numerical attribute, V_i its FP and $\mathcal{L}_i$ its upper delimiting path. Then $d^i_{*_1}(t,t')$ is defined as follows:*

$$d^i_{*_1}(t,t') = \begin{cases} 0 & \text{if } \exists v \in V_i, st.\ \mu_v(t) = \mu_v(t') = 1, \\ \frac{|\mathcal{L}_i(t,t')|}{|\mathcal{L}_i|} & \text{otherwise.} \end{cases} \tag{2}$$

Proposition 1. *The proposed definition of d^i_* when A_i is numerical is a dissimilarity.*

Proof. The dissimilarity between two values t and t' wrt. a strong FP being computed as the ratio between two path lengths, then the obtained dissimilarity degree is obviously positive and symmetrical. About the identity of indiscernibles, that should be interpreted in our case as the identity of indistinguishables, the conditional definition of $d^i_*(t,t')$ is used to guarantee such an indistinguishability relation between values inside the core of a fuzzy set. If $\mu_v(t) = \mu_v(t') = 1$ and due to the structural properties of the strong FP used on numerical attributes then $d^i_*(t,t') = 0 \Leftrightarrow R^{A_i}_t = R^{A_i}_{t'}$.

Remark 1. The satisfaction of the identity of indistinguishables is in opposition with the triangle inequality. Indeed, considering a partition element v and three points t, t' and t'' such that $\mu_v(t) = \mu_v(t') = 1$, $\mu_v(t'') < 1$ and $t \leq t' < t''$ (resp. $t'' < t \leq t'$), then $d^i_*(t,t') = 0$ and $d^i_*(t,t'') \geq d^i_*(t',t'')$ (resp. $d^i_*(t,t'') \leq d^i_*(t',t'')$). Thus, one observes that $d^i_*(t,t'') > d^i_*(t,t') + d^i_*(t',t'')$ which violates the triangle inequality property. The triangle inequality is however satisfied if there is no situation of indistinguishability between the three values considered.

Example 3. If one goes back to the situation depicted in Fig. 1 and Tab. 1, then the proposed definition of $d^i_*(t,t')$ leads to the expected behavior as shown by the dissimilarity matrix Tab 2.

3.2 For Categorical Attributes

Contrary to numerical attributes, categorical ones are generally defined on non-ordered domains. Hence, no explicit distance can be directly defined for a categorical attribute.

Table 2. Distance matrix between the items detailed in Tab. 1

	t_1	t_2	t_3	t_4	t_5	t_6
t_1	0	0.18	0.22	0.27	0.36	0.76
t_2	0.18	0	0.04	0.09	0.18	0.58
t_3	0.22	0.04	0	0	0.15	0.55
t_4	0.27	0.09	0	0	0.09	0.5
t_5	0.36	0.18	0.15	0.1	0	0.4
t_6	0.76	0.58	0.55	0.49	0.4	0

The question of computing a distance between categorical values has already largely been addressed, especially by the data mining community [9, 10]. Most of the proposed measures rely on contextual information, structural properties (considering clusters of data for instance) or correlations with other dimensions than the concerned categorical one [11, 12]. The seldom measures that only make use of the concerned categorical attribute deduce links between categorical values if their frequency of appearing is close [13, 14]. So, to the best of our knowledge, the dissimilarity defined in this section is the first one that addresses the question of comparing categorical values according to a discrete FP.

By defining a fuzzy-set-based vocabulary on a categorical attribute, the user expresses a subjective point-of-view about the way the categories have to be interpreted. A discrete fuzzy set gathers categories that define, combined all together, a "semantic concept". Categories regrouped in a same fuzzy partition element can be discriminated according to their respective membership degree within the set. When the user gradually assigns a categorical value to two different partition elements, we consider that he/she creates a semantic link between the two fuzzy-sets concerned. The idea behind the dissimilarity we propose for categorical attributes is to deduce, not an order, but semantical links between partition elements based on their intersections. These links are used in the proposed dissimilarity measure to compute a distance between two categorical values that belong to two different partition elements. The relevance of the links deduced between fuzzy sets based on their intersections obviously depends on the concerned applicative context and the semantics of point-of-view expressed by the user in his/her vocabulary.

Principles and Properties of the Proposed Dissimilarity Let t and t' be two categorical values satisfying the fuzzy terms v and v' respectively (v and v' may be identical) from an FP V_j. The principle of the dissimilarity measure is to combine the membership of t and t' to their respective partition elements (i.e. v and v') and the semantic closeness of v and v'. In other words, the more t and t' belong to a "semantically" close partition elements, the closer they are.

This semantic closeness between two elements from an FP is denoted by $C_J(v, v')$ and may be defined by means of the Jaccard index that quantifies the

proportion of elements v and v' share.

$$C_J(v, v') = \frac{\sum_{x \in \mathcal{D}} \min(\mu_v(x), \mu_{v'}(x))}{\sum_{x \in \mathcal{D}} \max(\mu_v(x), \mu_{v'}(x))}.$$

Definition 4. *For a categorical attribute A_i, the measure d^i_* is defined as follows:*

$$d^i_*(t, t') = 1 - \max_{j,k=1..q_i} \top(\mu_{v_{i,j}}(t), \mu_{v_{i,k}}(t'), C_J(v_{i,j}, v_{i,k})), \tag{3}$$

where the product t-norm $\top$ is used in our case for aggregating $\mu_{v_{i,j}}(t)$, $\mu_{v_{i,k}}(t')$ and $C_J(v_{i,j}, v_{i,k})$ to introduce compensation between the aggregated criteria.

Proposition 2. *d^i_* as defined in Eq. 3 is a dissimilarity.*

Proof. The definition of d^i_* when the concerned attribute A_i is categorical is obviously positive as both $\mu_{v_{i,j}}(t)$, $\mu_{v_{i,k}}(t')$ and $C_J(v, v')$ are defined in the unit interval.The Jaccard index and the product t-norm being symmetric, their combination in d^i_* is so as well. $d^i_*(t, t') = 0$ iff. $\mu_{v_{i,j}}(t) = 1$, $\mu_{v_{i,k}}(t') = 1$ and $C_J(v, v') = 1$. Due to the constraints imposed on the FP (Sec. 2.1) and especially the fact that each item is completely rewritten by $\mathcal{V}$ then $\mu_{v_{i,j}}(t) = 1$ (resp.$\mu_{v_{i,k}}(t') = 1$) and $C_J(v, v') = 1$ implies $v = v'$ and $R^i_t = R^i_{t'}$.

Remark 2. We consider that it would be artificial and senseless to introduce a notion of transitivity in the definition of d^i_*. It would indeed be debatable to consider that a value belonging to a partition element v_i is somewhat similar to a value belonging to an element v_j because v_i has a non-empty intersection with v_k that itself has a non-empty intersection with v_j, especially if v_i and v_j have an empty intersection.

Example 4. Tab 3 gives some dissimilarities computed between different car brands wrt. the FP illustrated in Fig. 2.

Table 3. Dissimilarity matrix for some car brands according to the FP Fig. 2

	VW	**Mercedes**	**AUDI**	**Ford**	**Peugeot**	**Daewoo**
VW	0	0	0.2	0.4	0.8	1
Mercedes	0	0	0.2	0.4	0.8	1
AUDI	0.2	0.2	0	0.4	0.8	1
Ford	0.4	0.4	0.4	0	0.8	1
Peugeot	0.8	0.8	0.8	0.8	0	0.75
Daewoo	1	1	1	1	0.75	0

4 Conclusion and Perspectives

The rewriting of data according to a fuzzy user vocabulary makes it possible to personalize a data-to-knowledge process. In order to be able to apply data

mining tools on linguistic and subjective representations of the data, it is first necessary to address the question of quantifying the dissimilarity between two such representations. We thus provide in this paper a dissimilarity measure that takes into account the structure of the fuzzy partitions that form the user vocabulary. We show on some examples that the proposed dissimilarities return relevant results and better discriminate the compared values without sacrificing the indistinguishability relation introduced by the use of fuzzy partition elements.

The next step is obviously to show that the use of this dissimilarity by a clustering algorithm leads to more meaningful and relevant results thanks to a better discrimination of the compared items.

References

1. Smits, G., Pivert, O., Yager, R.R.: A soft computing approach to agile business intelligence. In: Fuzzy Systems (FUZZ-IEEE), 2016 IEEE International Conference on, IEEE (2016) 1850–1857
2. Smits, G., Pivert, O.: Linguistic and graphical explanation of a cluster-based data structure. In: Scalable Uncertainty Management. Springer (2015) 186–200
3. Guillaume, S., Charnomordic, B., Loisel, P.: Fuzzy partitions: a way to integrate expert knowledge into distance calculations. Information sciences **245** (2013) 76–95
4. Smits, G., Pivert, O., Girault, T.: Reqflex: fuzzy queries for everyone. Proceedings of the VLDB Endowment **6**(12) (2013) 1206–1209
5. Ruspini, E.H.: A new approach to clustering. Information and Control **15**(1) (1969) 22 – 32
6. Bloch, I.: On fuzzy distances and their use in image processing under imprecision. Pattern Recognition **32**(11) (1999) 1873–1895
7. Montes, S., Couso, I., Gil, P., Bertoluzza, C.: Divergence measure between fuzzy sets. International Journal of Approximate Reasoning **30**(2) (2002) 91–105
8. Bloch, I.: Fuzzy geodesic distance in images. In: International Workshop on Fuzzy Logic in Artificial Intelligence, Springer (1995) 153–166
9. Boriah, S., Chandola, V., Kumar, V.: Similarity measures for categorical data: A comparative evaluation. In: Proceedings of the 2008 SIAM International Conference on Data Mining, SIAM (2008) 243–254
10. Alamuri, M., Surampudi, B.R., Negi, A.: A survey of distance/similarity measures for categorical data. In: Neural Networks (IJCNN), 2014 International Joint Conference on, IEEE (2014) 1907–1914
11. Gibson, D., Kleinberg, J., Raghavan, P.: Clustering categorical data: An approach based on dynamical systems. Databases **1** (1998) 75
12. Guha, S., Rastogi, R., Shim, K.: Rock: A robust clustering algorithm for categorical attributes. Information systems **25**(5) (2000) 345–366
13. Goodall, D.W.: A new similarity index based on probability. Biometrics (1966) 882–907
14. Lin, D., et al.: An information-theoretic definition of similarity. In: Icml. Volume 98. (1998) 296–304