On the Intuitiveness of Common Discretization Methods

[Short Version]

Mario Boley
Cluster of Excellence MMCI and Saarland
University Saarbrücken, Germany
mboley@mmci.uni-saarland.de

Ankit Kariryaa
University of Bielefeld
Bielefeld, Germany
ankit.ky@gmail.com

ABSTRACT

Data discretization methods are usually evaluated in terms of technical criteria that are related to some specific data analysis goal like the preservation of variable interactions. In this paper, we provide a different evaluation principle that assesses the quality of a chosen discretization as the degree to which it coincides with human intuition. This is motivated from the setting of interactive exploratory data analysis where discretizations should be simple, self-explanatory, and fix across results in order to reduce the cognitive load on the user. We present a study design for measuring the intuitive discretization choices of a general human population for a set of discretization problems and present the results of a study trial that we performed with 153 respondents and four problem classes—each using the categories "low", "normal", and "high". Through this trial, we evaluated eight discretization methods from three families: range-based discretization, count-based discretization, and clustering-based discretization. Our results partially confirm results from Cognitive Linguistics that assume prototype-based categorization, which is most closely resembled by clustering-based methods, as a predominant human discretization mechanism. They also show, however, an affinity of participants to sometimes compromise cluster quality in favor of approximating certain category proportions.

1. INTRODUCTION

Metric measurements, i.e., numerical data adhering to an interval or a ratio scale, are ubiquitous in real-world data analysis. Yet, many analysis algorithms require at least part of their input data in the form of simple binary features (e.g., Subgroup Discovery [Atzmueller, 2015], Redescription Mining [Parida and Ramakrishnan, 2005], and various data summarization techniques [Wille, 2005, Vreeken et al., 2011, Geerts et al., 2004]). This is why the data mining and statistics literature provides a wide range of data discretization techniques that can be used for producing such features from metric input (see, e.g., Kontkanen and Myllymäki [2007], Chapeau-Blondeau and Rousseau [2009], Nguyen et al. [2014]). Usually these techniques are evaluated solely from the technical perspective of how well they retain properties of the original data distribution and/or how they affect the performance of specific data analysis algorithms. In this paper we provide the complementing evaluation perspective of **intuitive linguistic discretization** in which one asks *how well is a discretization enabling an effective interaction between computer algorithms and human users as well as facilitating a discussion of algorithmic findings among humans.*

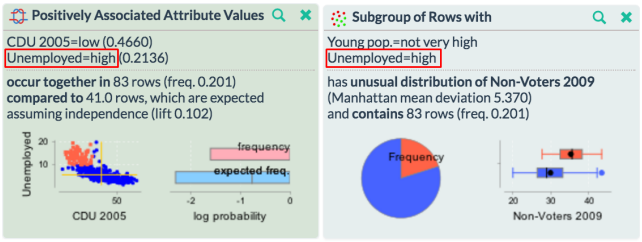

Figure 1: Two data analysis results produced by different algorithms in Creedo [Boley et al., 2015], both of which use the self-explanatory symbol "unemployed=high"; it is desirable that symbol has fix definition across results and that this definition is intuitive, i.e., approximately coinciding with user's category "high" (were she to know the distribution of "unemployment").

This perspective is relevant whenever algorithmic results are supposed to be interpreted by humans; especially when there are many such results as it is characteristic for exploratory data analysis and pattern discovery tasks. For example, consider a data scientist operating an interactive pattern discovery suite (e.g., MIME [Goethals et al., 2011], Cortana [Meeng and Knobbe, 2011], or VIKAMINE [Atzmueller and Lemmerich, 2012]). Typically, the scientist would run a number of data analysis algorithms with different parameter settings, the results of each of which she would investigate and compare with one another. Finally, she would distill out the most important findings for further discussion with her peers. From this scenario we can derive several desirable properties for discretization:

1. Since the results of different methods and different parameters should be comparable to one another, we want a stable and *generic* discretization that works

KDD 2016 Workshop on Interactive Data Exploration and Analytics (IDEA'16) August 14th, 2016, San Francisco, CA, USA.

ACM ISBN .

DOI:

reasonable well for various tasks and typical analysis methods. This is in contrast to discretizations that are optimized for one specific setting as it is the case for supervised discretization techniques.

2. Moreover, the discretization should be *self-explanatory* in order to reduce the cognitive load of the data scientist. That is, we are looking for a discretization that summarizes metric variables in a comprehensible way through a small number of linguistic terms like "low", "normal", and "high".

3. Finally, the discrete symbols should be *intuitive*. That is, ideally the symbols' definitions approximately correspond to those that humans would instinctively pick themselves to talk about the data domain among each other.

Fig. 1 summarizes these criteria for an exemplary result set produced by different pattern discovery algorithms. Based on requirements 1 and 2, we think of an abstract (exact) **linguistic discretization problem** as: *given* a sample S of values of a metric variable defined on a real interval X and a set of k ordered linguistic quantification categories, *find* $k-1$ cut-off values in X that separate the given categories for that variable. Based on requirement 3, we say that a discretization given by a set of cut-off values is **intuitive** if it tends to be close to the set of cut-off values that users of a desired target audience would pick themselves had they to make their choice purely based on the sample S (as opposed to *concrete* linguistic discretization tasks where prior information about the variable is available). In this article we investigate empirically the degree to which common discretization approaches exhibit this form of intuitiveness.

Studying the precise mechanism of human discretization is a profound topic with connections to Linguistics (where it is referred to as *categorization*, see Taylor [2003] and references therein) as well as Cognition and Neuroscience (e.g., Dehaene et al. [1998, 2008]). Here, we generally take on a rather naive point of view and simply propose to test how well algorithmic discretization of quantities aligns with human categorization while staying agnostic about the precise mechanism that governs it. A particular interesting proposition from Cognitive Linguistics [Evans, 2007], that we take up here, is that the predominant mechanism for human linguistic discretization is based on prototypes (going back to a seminal work of Rosch [1973] in Cognitive Psychology). This proposition says that categories are associated with typical representative members (and that there can be values that are not a real representative of any category). In Computer Science this intuition was formalized as fuzzy linguistic discretization through fuzzy logic (see Ishibuchi et al. [2006] and references therein). This approach, however, requires specific analysis and model induction algorithms. Here we are interested in a general purpose preprocessing method, and, hence, we focus on traditional interval-based (or exact) discretization methods. Among those, clustering-based methods come closest to the idea of prototype-based categorization. Therefore we put a special emphasis on the evaluation of those methods.

To summarize the **contributions** of this paper: firstly, we develop a study design for measuring the intuitive discretization choices of a general target audience and that therefore operationalizes all of the theoretical concepts mentioned above. Secondly, we report results that have been generated with this design through an open study trial involving 153 participants that was particularly targeting the general categories "low", "normal", and "high". Our findings partially confirm the prototype-based proposition, but also show that it is violated when the distribution of the input sample is spread out too uniformly. In particular, we observed an affinity of participants to sometimes compromise cluster quality in favor of approximating certain category proportions.

Figure 2: Cut-off values of geometric-width discretization for $k = 7$ and $X = [0, 1]$ or the quantiles of geometric-frequency labeling for $k = 7$.

2. FORMAL DISCRETIZATION METHODS

In this section we define the formal discretization methods that we want to evaluate. First, however, we need to fix some basic notation. Let $X = [a, b] \subseteq \mathbb{R}$ be the real interval given by the upper and lower bounds $a, b \in \mathbb{R}$, respectively. We are interested in categorizing elements of X into a fixed number k of ordered discrete categories $K = \{1, \ldots, k\}$. To a human user these categories would be presented as interpretable words like $\{\text{extremely low}, \text{very low}, \ldots, \text{extremely high}\}$. A **discretization** of X is a function $c : X \to \{1, \ldots, k\}$ given by $k-1$ cut-off values $c_1 < c_2 < \ldots, c_{k-1}$ through $c(x) = \min\{i : c_i \geq x\}$. An (empirical) **discretization method** maps finite samples $S \subseteq X$ to a uniquely defined discretization. As a convention we define as $S = \{s_1, \ldots, s_n\}$ with $s_i \leq s_j$ for $i < j$. Many discretization methods actually only yield a **labeling**[1] $l : S \to K$ of the given sample rather than cut-off values on the real interval. For those cases, we consider the **canonical discretization** of a labeling l as the one given by the cut-off values

$$c_i = (\max\{s \in S : l(s) = i\} + \min\{s \in S : l(s) = i+1\})/2 \ ,$$

for $i \in \{1, \ldots, k-1\}$. That is, cut-off values are defined as the arithmetic mean between the extreme values of adjacent category labels.

The first and most simple family of discretization methods that we consider are **ranged-based methods** that define cut-off values as a simple function of the underlying interval (sample-independent variant) or the range of the given data sample (sample-dependent variant). The simplest member of this family is **sample-independent equal-width discretization**, which is given by the cut-off values

$$c_i = a + i(b-a)/k$$

for $i \in \{1, \ldots, k-1\}$. For **sample-dependent equal-width discretization** the smallest and the largest sample element are used in place of the interval boundaries a and b, i.e., cut-off value i is defined as $s_1 + i(s_n - s_1)/k$. While these methods are very simple to define, depending on the given category names, both of these approaches

[1]Labelings resulting from discretization methods of course must be monotone, i.e., $l(s) \leq l(s')$ if $s \leq s'$.

(a)

(b)

Figure 3: Example populations of cups (a) and sunglasses (b) for the *prize* narrative in the study trial.

can be counter-intuitive: for example for "high", "normal", and "low" they set the normal range to be of equal size as the two extreme ranges. Therefore, for an odd number of categories $k > 2$, we define sample-dependent and sample-independent **geometric-width discretization** as alternative range-based approaches that cut the range into increasingly fine pieces when approaching the interval (or sample) borders. That is, for the sample-independent variant, the cut-off values are defined as

$$c_i = \begin{cases} b - (b-a)g_{(k-1)/2-i+2}, & \text{for } i \leq k/2 \\ a + g_{i-(k-1)/2+1}, & \text{for } k/2 < i < k \\ 1, & \text{for } i = k \end{cases}$$

with the geometric sums $g_m = \sum_{j=1}^{m} 2^{-j}$, and for the sample-independent variant, a and b are again replaced by s_1 and s_n, respectively. See Fig. 2 for an illustration.

As a second family of discretization methods we consider **frequency-based discretizations**. Those methods determine labelings based on desired counts of data values per category and are indifferent to the metric proximity between values. Technically, these labelings are most conveniently defined through the sample **quantiles** $q(\alpha) = \min\{s_i \in S : i/n \geq \alpha\}$ for $\alpha \in [0, 1]$. A sequence of fractions $\alpha_1 < \alpha_2 < \cdots < \alpha_k = 1$ gives rise to a labeling $l(s) = \min\{i \in K : p(\alpha_i) \geq s\}$. The most well-known instantiation of this scheme is **equal-frequency labeling**, which uses the set of equidistant quantiles given by $\alpha_i = i/k$ for $i \in K$. Again it can be linguistically somewhat counter-intuitive when all categories contain an equal number of sample values. For "low", "normal", and "high", this would imply that only a minority of data-values is considered "normal" and two third are either "high" or "low". To address this issues, for odd $k > 2$ we again define a variant based on increasingly refined categories (this time in terms of the quantiles), that we refer to here as **geometric-frequency labeling**. It is given by the fractions

$$\alpha_i = \begin{cases} 1 - g_{(k-1)/2-i+2}, & \text{for } i \leq k/2 \\ g_{i-(k-1)/2+1}, & \text{for } k/2 < i < k \\ 1, & \text{for } i = k \end{cases}$$

where g_m denotes the geometric sum as above.

As a final family of discretization methods we consider **clustering-based methods**. These methods determine a labeling based on a set of k reference values $R \subset X$, each of which is the representative for one of the categories. Assuming that R consists of the elements $r_1 < r_2 < \cdots < r_k$, the resulting labeling is then defined by $l(s) = i$ where r_i is a reference value that minimizes $|s - r|$ with $r \in R$ (breaking ties, e.g., by using the minimal such value). Naturally, one wants to use the set of reference values that are closest to their associated sample values. If one uses the sum of squared differences, $\sum_{s \in S}(r_{l(s)} - s)^2$, to measure this closeness, this approach yields **k-means-based labeling** (the mean of a set of values minimizes the sum of the squared distances). Since the reference values in this approach can be arbitrary elements of the underlying interval X they are susceptible to outlying sample values, which can lead to counter-intuitive discretizations. This can be addressed by using the reference values that minimize the sum of absolute errors, $\sum_{s \in S} |r_{l(s)} - s|$. Since, the sum of absolute errors of a set of values is minimized by any median value of that set, this variant is called **k-medians-based labeling**.

3. EMPIRICAL DESIGN

In this section we develop the study design (empirical method) for comparing formal discretization methods to discretization performed by humans. This includes a questionnaire for posing abstract discretization tasks to a general audience, a set of discretization tasks as re-usable test cases for the given as well as for follow-up studies, and measures for the quantification of the similarity of human and formal discretization results.

3.1 Questionnaire

The purpose of the questionnaire is to gather data from members of a general target audience on how they intuitively perform abstract linguistic discretization tasks. This measurement problem entails a central difficulty. While potential participants are used to perform intuitive discretization for concrete variables, it is likely to not work as intended to directly pose to them an abstract task about an unknown variable: it would trigger a formal approach to the problem and/or possibly yield a low engagement with the task and consequently relatively arbitrary answers.

Therefore, the key idea of our questionnaire design is to decorate the abstract tasks with concrete **narratives** of tangible variables from everyday life. The trick is that we use variables that have a value distribution which greatly depends on the specific sub-population they are defined on, and then to leave the sub-population ambiguous—with the given sample as the only means to infer it. This way the task has to be solved factually with the same information as the underlying abstract task. Of course, the given narrative might still influence the responses. It is therefore advisable to use multiple narratives so that their effects cancel out when averages over the whole result set are taken.

In our study trial, we opted for two narratives: *prices of products* and *ages of humans*. For both variables, one is used to heavily alter the usage of quantification terms across different classes of products and groups of people, respectively. For instance, even a "very low" price for a TV set is likely to be considered "high" if it were the price of a light bulb. Similarly, the age of a "young" high-school teacher would be considered "older" for a college student. The questionnaire design emphasizes this sub-population dependency by introducing the narrative with two named and labeled **example populations** that show a contrasting variable distribution. In our study trial, we used for the prize narrative the examples of cups and sunglasses (see Fig. 3) and for the age narrative the examples of members of a fencing team and

In different contexts numbers can have different interpretations. What you consider [**example category 1**] in one case, you might consider [**example category 2**], or [**example category 3**] in another. Consider the image below of a set of [**example population 1**]. The number underneath each [**population member**] shows [**variable**] in [**unit**]. The labels below the picture show an exemplary categorization of the numbers into [**category list**] that you perhaps would roughly agree to in the context of this set.

[**image of population 1 with variable labels**]
[**example cut-off points 1**]

Now compare this categorization into [**category list**] to the next categorization for [**variable**] of [**example population 2**]. Despite being different, each of the categorizations make sense in their respective set.

[**image of population 2 with variable labels**]
[**example cut-off points 2**]

Text 1: Leading text of questionnaire, which introduces narrative along with example populations.

the inhabitants of an elderly housing facility. In the questionnaire, images of the example populations are embedded into an introductory text that explains the sub-population dependence of the linguistic terms. The verbatim text-frame is given in Text 1.

Following this introductory passage, a number of actual discretization tasks is presented to the participant. In order to support the narrative, the sample values are embedded into images that depict anonymous populations for the variable. For the two narratives in our trial those images are given in Figs. 4 and 5, respectively. The tasks are introduced with the instruction text given in Text 2. Note that we do explicitly mention the possibility of choosing cut-off values that are not part of the given sample itself. This possibility can be further emphasized by using this option in the example discretizations.

In summary the proposed questionnaire design allows to pose a number of abstract linguistic discretization tasks to participants from a general population by decorating them with a concrete narrative. It is required that all tasks on one instance of the questionnaire use the same linguistic categories and that their sample ranges match the chosen narrative. This might require to rescale some of them. In the next subsection, we discuss these and other issues abounding when creating a full study design around this questionnaire.

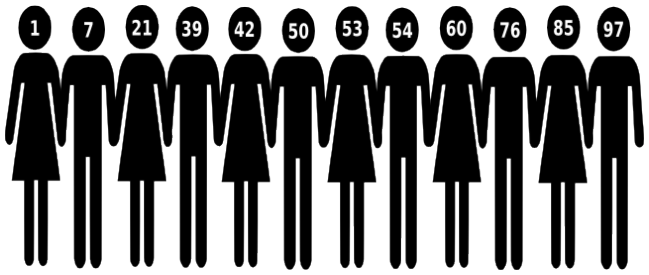

Figure 4: Task image for *age* narrative.

In this short survey, we ask for your opinion on what means [**category list**] in the context of three anonymous groups of [**population type**]. Underneath each of the images below, please fill into the designated boxes what you consider [**category list**]. Note that you can fill in numbers that do not occur in the sample itself. When you are done, please do not forget to click the submit button. Thanks a lot for your participation.

[**image of task 1**]
[**input fields for cut-off values**]
...
[**image of task z**]
[**input fields for cut-off values**]

Text 2: Instructions and task part of questionnaire.

3.2 Discretization tasks

When setting the discretization tasks for the study, there are two components that have to be defined more or less independently: the linguistic categories to be used as well as the actual numerical samples. Regarding the first component it is important to note that the validity of any results of the study, when interpreted strictly, is tied to the specific quantifiers used. Although certain insights can arguably be transferred between different category sets, it is generally unclear whether the human expectation for appropriate interval sizes varies depending on if they are called "low", "normal", and "high" or, e.g., "reduced", "moderate", and "increased". Similarly, quantifiers for specific kinds of variables, e.g., "long" for length, might carry their own expectational bias and may be not fully compatible to their generic counter-parts.

In the given instantiation of the study design we choose to focus only on categorization into

$$K = \{\text{“low”}, \text{“normal”}, \text{“high”}\} \ .$$

The rational for this choice was that these are the perhaps most widely applicable quantifiers for numerical values. Moreover, using three categories arguably constitutes a pareto-optimal choice when trading off the interpretability of the categories (individually and jointly) and their accuracy in representing the underlying numerical range.

Turning to the samples, the goal is to have a diverse set of tasks which is likely to allow to differentiate between the different discretization methods even with a relatively small number of values. On the one hand, for the aim to have a consistently high response quality it is desirable to work with small samples. The larger the sample size the more variation is likely to occur among participants in the degree to which they fully process individual sample values. On the other hand, the smaller the sample the lesser the results are likely to generalize to realistic sample sizes in Data Analysis. In particular, seven plus/minus two apparently constitutes a phase transition between the usage of different mental processing mechanisms according to the classic result of Miller [1956]. Therefore, in the given study we use the **sample size** $|S| = 12$ throughout all discretization tasks. Moreover,

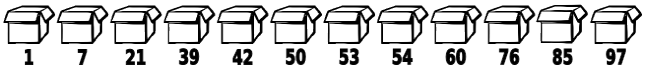

Figure 5: Task image for *prize* narrative.

Task	Class	Sample	c_1	c_2
a)	Uniform	$\{5, 18, 24\}$, $\{30, 32, 32, 50\}$, $\{70, 75, 87, 91, 95\}$	25	70
b)	Normal	$\{1, 7, 21\}$, $\{39, 42, 50, 53, 54, 60\}$, $\{76, 85, 97\}$	25	65
c)	Exponential	$\{40, 42, 42, 43\}$, $\{47, 47, 48, 48, 53\}$, $\{62, 70, 74\}$	45	60
d)	Mix	$\{4, 6, 8, 10\}$, $\{22, 23, 24, 25, 28\}$, $\{37, 49, 56\}$	11	35
e)	Uniform	$\{15, 27, 27\}$, $\{\underline{31}, \underline{35}, \underline{37}, 51, 53, 54\}$, $\{80, 81, 90\}$	30	70
f)	Normal	$\{31, 31, 35, 39\}$, $\{77, 79, 82, 82, 82\}$, $\{93, 93, 98\}$	40	90
g)	Exponential	$\{24, 24, \underline{34}\}$, $\{41, 43, 46, 49, \underline{56}\}$, $\{63, 64, 65, 81\}$	35	63
h)	Mix	$\{1, 3, 7\}$, $\{\underline{20}, 30, 37, 37, 38, 39, 44\}$, $\{55, 68\}$	18	46

Table 1: First two group of samples generated for task classes in study trial with median cut-off values of respondents. Underlined sample elements show discrepancy of resulting labeling with k-medoids.

in order to fit our narratives, we scale the **variable range** to $X = [0, 100]$ but, again in order to reduce the cognitive burden of the participants (and thus reduce variation in result quality) we work with rounded samples $S \in \{1, \ldots, 100\}^*$. In particular, we remove 0 from the sampling range in order to maintain the intuition of the price narrative.

For generating the samples, we define four **classes of discretization tasks**—uniform, normal, exponential, and mixture—based on four continuous random variables with probability density functions p_{uni}, p_{nrm}, p_{exp}, and p_{mxt}, respectively. A discretization task for a class with pdf p is then generated by drawing a sample of 12 independent realizations of the rounded and truncated version of the corresponding random variable, i.e., using the probability mass function

$$f(n) = \int_{n-1}^{n} p(x)/Z \, dx$$

for $n \in \{1, \ldots, 100\}$ with $Z = \int_0^{100} p(x) \, dx$. The formal definitions of the continuous pdfs are as follows.

Uniform is simply defined by the uniform pdf $p_{\text{uni}}(x) = 1$ for $x \in [0, 100]$. For a task from this class we do not expect significant value clusters to appear. Hence, there might be a tendency among participants to resort to simpler principles than clustering-based discretization.

Normal is defined through $p_{\text{nrm}}(x) = \phi_{M,S}(x)$, i.e., the Gaussian pdf with uniform random mean M and standard deviation S drawn from $[0, 100]$. Tasks from this class are likely to show a central tendency towards a random mean. Hence, it can be expected that human assignment of "normal" will reflect that tendency in contrast to the sample independent range-based discretization methods.

Exponential is defined by $p(x) = O + R \exp(-Rx)$ with a uniform random offset term O from $[0, 50]$ and a uniform random rate parameter R from $[0.2, 0.8]$. Tasks from this class are expected to have a highly skewed distribution, which should render symmetric range-based discretizations counter-intuitive.

Mixture is defined by $p(x) = \phi_{M_1,S_1}(x) + \phi_{M_2,S_2}(x)$ as the mixture of two Gaussians with uniform random means and standard deviations as defined for the class normal. Samples from this class are expected to be bi-modal with a high, a low, and normal range around each mode. This is generally hard to reflect adequately with three categories only, but it is to be expected that count-based and clustering-based approaches can find reasonably intuitive compromises.

3.3 Evaluation measures

After designing the test discretization tasks as well as a questionnaire for querying human solutions to these tasks, it remains to define how we want to compare the discretizations produced by formal methods with those of the study participants. We will do this on two levels of resolution: on the first, we just compare the discretizations in terms of how they label the given sample; on the second, we measure the difference of the actual cut-off points.

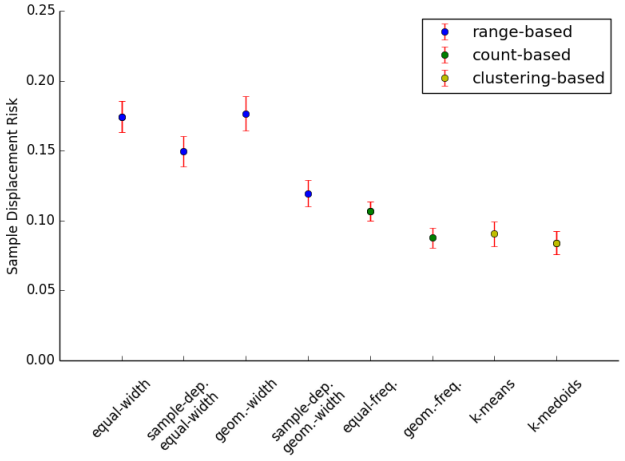

Figure 6: Sample displacement risk per method over all tasks in study with 95%-confidence intervals.

Let c and d be two discretizations of the range $X = [a, b]$ using the categories $K = \{1, \ldots, k\}$ and $S \in X^*$ be a finite sample of X. Independent of whether we want to quantify the difference between c and d through their cut-off values on the whole range X or just in terms of how they label the elements of S, we first have to fix the **displacement loss** between two categories, i.e., how much we consider it harmful to use a category j in place of the true category i. For that we propose to use the relative difference of the category numbers $l(i, j) = |i - j|/(k - 1)$ (we normalize here with $k - 1$ rather than k so that l reduces to the 0/1-loss when $k = 2$).

When evaluated on all category pairs abounding from applying c and d to the sample S, this loss function leads to the **sample displacement loss** for discretizations defined by

$$l_S(c, d) = \frac{1}{|S|} \sum_{x \in S} l(c(x), d(x)) \ .$$

As described above, this measure considers the categoriza-

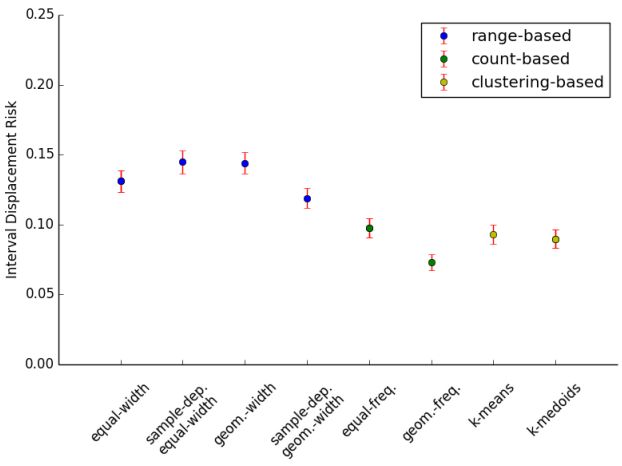

Figure 7: Interval displacement risk per method over all tasks in study with 95%-confidence intervals.

tions as mere labelings of the data sample and, beyond that, does not reflect how the discretizations differ when viewed as linguistic categorizations of the underlying domain. For that purpose we define the **interval displacement loss**

$$l_X(c,d) = \frac{1}{b-a}\int_{x\in X} l(c(x),d(x))\,dx$$

which quantifies disagreement between the discretizations in terms of the size of the underlying domain-pieces with a certain displacement. This integral can be simply computed as the sum of the piece-wise constant losses on the intervals resulting from cutting X with all cut-off values in c and d.

Now assume we have a bag of study results R consisting of pairs of samples and discretizations $\{(S_1,d_1),\ldots,(S_m,d_m)\}$ where discretization d_i is the result of a respondent for the task involving sample S_i. For a discretization method $m : X^* \to (X \to K)$ we can then determine its **empirical sample displacement risk** as

$$r_R = \frac{1}{R}\sum_{(S,d)\in R} l_S(m(S),d)\ .$$

Given that the set of respondents and the set of tasks is representative for some larger population of tasks and users of interest, this empirical risk will approximate the real population displacement risk (defined through the expected loss-value over this population) for the method m. Switching to the interval displacement loss, we can define similarly the empirical interval displacement risk based on a set of study results as well as the underlying population displacement risk.

4. RESULTS

In this section we report the results of an open online study trial[2] using the design developed in Sec. 3. The trial was conducted over the course of 10 days with a total of 153 respondents. We advertised the study through a call for participation that was published via internal mailing lists of 6 academic institutions from 4 different countries (UK, Germany, Finland, and Israel) as well as through social media in 3 different networks (Facebook, LinkedIn, and Google+). The call encouraged participants to re-share the invitation for participation with potentially interested colleagues and friends. Hence it was a convenience- and snow-ball sampling scheme of participants with the goal to maximize the number of responses—sacrificing control over the participant demographics.

For the same purpose and also to ensure an as high as possible quality of responses, we intended to keep the expected time and attention for participation low. Hence, we settled to issue only 3 discretization tasks per participant. Moreover, since this was the first study of this kind, we wanted to be able to meaningfully inspect the results, in particular, the chosen cut-off values for each of the generated tasks. For that reason we opted for a task sampling scheme that generates a certain number of repetitions per sample (for each narrative). In more detail, we iteratively fixed groups of 4 random tasks (one per task class). Three tasks of a group were then issued to each requested questionnaire uniformly at random (decorated by a random narrative) until each task in the group (for each narrative) received at least 25 responses. Only then a new group was generated. Thus, we traded off representativeness for the individual task classes for an attempt to acquire confident estimates of preferred cut-off values per sample. See Tab. 1 for a list of all samples for which the full number of responses was reached along with the median respondents' cut-off values.

4.1 Overall Outcome

Aggregating over all trial results, i.e, all responses for all discretization problems, the following picture emerges for the empirical sample displacement risk (see Fig. 6 and also Tab. 2, upper portion, row "all"). There are three methods leading the field: k-medoids-based labeling (risk[3] of apprixmately 0.0839 ± 0.0083), geometric-frequency labeling (0.0876 ± 0.0073), and k-means-based labeling (0.0904 ± 0.0089). While the results do suffice to confidently separate this group from the rest of the methods, they are insufficient to confidently separate them from one another. The next group consists of equal-frequency labeling (0.1067 ± 0.0069) and sample-dependent geometric-width labeling (0.1193 ± 0.0095). The remaining range-based methods are at the end of the spectrum with a small but significant advantage for sample-dependent equal-width (0.1494 ± 0.0109).

Turning to the interval displacement risk (see Fig. 7 and Tab. 2, lower portion, row "all"), the first observation is that the magnitude of empirical loss values is somewhat and their variation is notable smaller as for the sample displacement risk. Consequently we have smaller confidence intervals. The ranking of the methods are slightly shifted with geometric-frequency labeling (0.0729 ± 0.0058) now leading confidently in front of the following group consisting of k-medoids-based labeling (0.0898 ± 0.0063), k-means-based labeling (0.0931 ± 0.007), and labeling based on equal-frequency (0.0974 ± 0.0067). At the end of the field we have again the range-based methods. Out of those methods, just as with the sample displacement risk, sample-dependent geometric-width performs best. However, for the interval displacement

[2]All results can be downloaded from http://www.realkd.org/wp-content/uploads/2016/05/discretization-study-results.csv

[3]We give all risks here rounded to 4 digits with $\alpha = 0.95$ two-sided confidence intervals.

	si equal-width	sd equal-width	si geom.-width	sd geom.-width	equal-freq.	geom.-freq.	k-means	k-medoids
				sample displacement risk				
normal	0.1439 ± .0225	0.1150 ± .0188	0.1791 ± .0282	0.1150 ± .0188	0.1129 ± .0137	0.0916 ± .0177	**0.0789 ± .0185**	**0.0789 ± .0185**
exponential	0.1684 ± .0214	0.2419 ± .0168	0.2097 ± .0252	0.2216 ± .0152	0.1201 ± .0177	0.1019 ± .0156	**0.0875 ± .0169**	0.1078 ± .0154
uniform	0.1342 ± .0151	0.1988 ± .0170	**0.0841 ± .0129**	0.0844 ± .0121	0.1131 ± .0108	0.0910 ± .0126	0.1571 ± .0160	0.1074 ± .0161
mixture	0.2515 ± .0213	0.0413 ± .0124	0.2279 ± .0181	0.0498 ± .0111	0.0796 ± .0093	0.0649 ± .0091	0.0413 ± .0124	**0.0410 ± .0124**
all	0.1742 ± .0110	0.1494 ± .0109	0.1765 ± .0123	0.1193 ± .0095	0.1066 ± .0069	0.0876 ± .0073	0.0904 ± .0089	**0.0839 ± .0083**
				interval displacement risk				
normal	0.1254 ± .0125	0.1788 ± .0184	0.1240 ± .0132	0.1509 ± .0161	0.1410 ± .0162	**0.0898 ± .0147**	0.1160 ± .0168	0.1160 ± .0168
exponential	0.0792 ± .0102	0.1927 ± .0145	0.1427 ± .0126	0.1719 ± .0099	0.0730 ± .0114	0.0559 ± .0103	**0.0557 ± .0105**	0.0718 ± .0093
uniform	0.0941 ± .0110	0.1340 ± .0113	**0.0800 ± .0113**	0.0892 ± .0108	0.0954 ± .0121	0.0876 ± .0119	0.1420 ± .0114	0.1104 ± .0106
mixture	0.2272 ± .0110	0.0680 ± .0073	0.2280 ± .0108	0.0576 ± .0066	0.0782 ± .0080	**0.0585 ± .0061**	0.0607 ± .0081	0.0607 ± .0081
all	0.1310 ± .0077	0.1448 ± .0082	0.1439 ± .0078	0.1189 ± .0072	0.0974 ± .0067	**0.0729 ± .0058**	0.0930 ± .0070	0.0898 ± .0063

Table 2: Empirical sample displacement and interval displacement risks with 95% confidence intervals—taken over all tasks and per task class. All numbers are rounded to 4th digit after decimal point.

risk, its confidence interval has a slight overlap with the sample-independent equal-width method.

4.2 Outcome per task class

When looking at the results per task class (see Tab. 2), one can make some notable observations specifically when looking at the performance of the clustering-based methods. Both, k-means and k-medoids, are the best or among the best methods in all task classes but *uniform*. Here, k-means is the second worst with respect to sample displacement risk and the worst with respect to interval displacement risk. Notably, k-medoids performs more robust for this task class, while its ranks (4 and 6, respectively) also deviate substantially from the ranks it achieves for the other classes. In contrast to clustering-based methods, range-based discretization in the form of sample-independent and dependent geometric-width, are specifically strong for the uniform tasks. They are also competitive for mixture but perform both weakly for normal as well exponential.

Finally, we can observe that geometric-frequency performs consistently well across all problem types independent of the risk functional. In fact, for the interval displacement risk it performs best or at least not significantly worse than the best for all classes. Interestingly, when looking at the median cut-off values for all individual samples for which a large number of responses was generated (Tab. 1), we can see that there is only one sample (uniform sample 'k') for which the labeling of the median respondents' cut-off points disagrees substantially with k-medoids and two more (samples 'g' and 'h') where there is a minor disagreement. In all these cases, the respondents' median labeling has category frequencies closer to the geometric frequencies (0.25, 0.5, 0.25) than the solution of k-medoids (there is a similar trend for 'a' and 'g', where the median exactly respects the k-medoids objective, but there is still a substantial number of respondents who did not adhere to it and produced category proportions closer to the geometric frequencies).

5. CONCLUSIONS AND OUTLOOK

With the presented study design we were able to gather for the first time insights on the human expectation for discretizing numerical data into the discrete categories: "low", "normal", and "high", which are important categories, e.g., for providing a simple intuitive discretization in data analysis suites. Particular findings are:

1. Clustering-based methods appear to yield good results essentially confirming the proposition from Cognitive Psychology and Cognitive Linguistics, which says that humans tend to perform categorization based on similarity to category prototypes. It seems, however, that this mechanism alone is not enough to fully replicate human discretization choices for quantitative linguistic categories (such as "low", "normal", and "high'). In our trial we could observe a tendency to sometimes deviate from optimal clustering-based solutions, presumably, in order to create more satisfying category frequencies.

2. Particularly, for the chosen linguistic categories, a frequency of 0.5 for the "normal" category, and a frequency of 0.25 for the categories "low" and "high" each appear to be attractive. This is testified by the fact that the frequency-based method with these parameters performed robustly well across different tasks.

3. Ranged-based methods that disregard the sample (or almost disregard it except for the extreme values) appear to be too simplistic for robustly creating an intuitive labeling and can not compete with the other method classes. Hence, while the relative differences of metric attributes seem to have a notable effect on labeling decisions, the absolute scale of metric information seems to be at most of minor importance.

Generally, we hope that the given work will open the door for systematically deriving novel approaches for intuitive discretization and evaluating them with the proposed or a modified study design. Some open questions we consider to be of particular importance are the following.

1. To what degree are the trends we discovered representative for the underlying problem classes and for specific target audiences. In the performed trial, representativeness for task classes has been sacrificed for more representativeness of the individual tasks, and the population was mostly uncontrolled.

2. What is an intuitive mechanism for deriving precise cut-off points from a given labelling? The given trial data showed no clear trend of how human cut-off values relate to their labelings.

3. Finally, the perhaps most interesting direction for future research is to investigate to what degree the identified trends hold up for finer categorizations, e.g., into "very low", "low", etc. and larger samples per task. A particular question for the frequency-based methods is whether the geometric proportions are really the expected continuation of the 0.25/0.5/0.25 scheme.

References

Martin Atzmueller. Subgroup discovery. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 5(1):35–49, 2015.

Martin Atzmueller and Florian Lemmerich. Vikamine–open-source subgroup discovery, pattern mining, and analytics. In *Machine Learning and Knowledge Discovery in Databases*, pages 842–845. Springer, 2012.

Mario Boley, Maike Krause-Traudes, Bo Kang, and Björn Jacobs. Creedo—scalable and repeatable extrinsic evaluation for pattern discovery systems by online user studies. In *IDEA 2015*, 2015.

François Chapeau-Blondeau and David Rousseau. The minimum description length principle for probability density estimation by regular histograms. *Physica A: Statistical Mechanics and its Applications*, 388(18):3969–3984, 2009.

Stanislas Dehaene, Ghislaine Dehaene-Lambertz, and Laurent Cohen. Abstract representations of numbers in the animal and human brain. *Trends in neurosciences*, 21(8): 355–361, 1998.

Stanislas Dehaene, Véronique Izard, Elizabeth Spelke, and Pierre Pica. Log or linear? distinct intuitions of the number scale in western and amazonian indigene cultures. *Science*, 320(5880):1217–1220, 2008.

Vyvyan Evans. *A glossary of cognitive linguistics*, volume 251. Edinburgh University Press, 2007.

Floris Geerts, Bart Goethals, and Taneli Mielikäinen. Tiling databases. In *Discovery science*, pages 278–289. Springer, 2004.

Bart Goethals, Sandy Moens, and Jilles Vreeken. Mime: a framework for interactive visual pattern mining. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 757–760. ACM, 2011.

Hisao Ishibuchi, Tomoharu Nakashima, and Manabu Nii. *Classification and modeling with linguistic information granules: Advanced approaches to linguistic Data Mining*. Springer Science & Business Media, 2006.

Petri Kontkanen and Petri Myllymäki. Mdl histogram density estimation. In *International Conference on Artificial Intelligence and Statistics*, pages 219–226, 2007.

Marvin Meeng and Arno Knobbe. Flexible enrichment with cortana–software demo. In *Proc. Benelearn*, pages 117–119, 2011.

George A Miller. The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychological review*, 63(2):81, 1956.

Hoang-Vu Nguyen, Emmanuel Müller, Jilles Vreeken, and Klemens Böhm. Unsupervised interaction-preserving discretization of multivariate data. *Data Mining and Knowledge Discovery*, 28(5-6):1366–1397, 2014.

Laxmi Parida and Naren Ramakrishnan. Redescription mining: Structure theory and algorithms. In *AAAI*, volume 5, pages 837–844, 2005.

Eleanor H Rosch. Natural categories. *Cognitive psychology*, 4(3):328–350, 1973.

John R Taylor. *Linguistic categorization*. OUP Oxford, 2003.

Jilles Vreeken, Matthijs Van Leeuwen, and Arno Siebes. Krimp: mining itemsets that compress. *Data Mining and Knowledge Discovery*, 23(1):169–214, 2011.

Rudolf Wille. Formal concept analysis as mathematical theory of concepts and concept hierarchies. In *Formal Concept Analysis*, pages 1–33. Springer, 2005.