Towards a User-Defined Visual-Interactive Definition of Similarity Functions for Mixed Data

Jürgen Bernard*
Fraunhofer IGD, Germany
Marco Hutter§
Fraunhofer IGD, Germany
David Sessler†
Fraunhofer IGD, Germany
Tobias Schreck¶
University of Konstanz, Germany
Michael Behrisch‡
University of Konstanz, Germany
Jörn Kohlhammer‖
Fraunhofer IGD, Germany
TU Darmstadt, Germany

Abstract

The creation of similarity functions based on visual-interactive user feedback is a promising means to capture the mental similarity notion in the heads of domain experts. In particular, concepts exist where users arrange multivariate data objects on a 2D data landscape in order to learn new similarity functions. While systems that incorporate numerical data attributes have been presented in the past, the remaining overall goal may be to develop systems also for mixed data sets. In this work, we present a feedback model for categorical data which can be used alongside of numerical feedback models in future.

Index Terms: I.5.3 [Computing Methodologies]: Pattern Recognition—Clustering; H.5.2 [Information Systems]: Information Interfaces and Presentation—User Interfaces

1 Introduction and Previous Work

Data sets containing categorical and numerical data attributes (mixed data) occur in practically all application domains. Here, typical analytical tasks like searching for nearest neighbors, grouping similar objects, detecting outliers or recognizing other interesting patterns can only be conducted if a similarity function is provided. Such a function computes distance scores between multivariate data objects, respectively. However, individual objects of these mixed data sets are hard to compare. This is especially the case if the user's mental similarity notion (short: MSN) is subject to change over time, as typically seen in sense-making loops. In these cases a *user-guided similarity definition* allows to apply and adapt the MSN of domain experts interactively, at run-time. As an example, the user arranges the objects Berlin, Paris, and Washington close to each other on a 2D landscape, and the system learns that the MSN of the user concerns the categorical attribute 'Capital City'. Another, more academic example might be a 2D landscape of patient data, where a doctor arranges patients close to each other who all had a typical sort of behavior (based on the MSN of the expert). The beneficial means of such a system would be that attributes would protrude for an early detection/diagnosis, possibly undiscovered as such so far. This utilization of *user-defined object arrangement* is already applied for numerical attributes in inspiring works of Liu et al. [2] and Mamani et al. [3]. However, the development of *feedback models* which also cope with mixed data remains challenging.

In a previous work, we presented a concept for the development of systems for user-defined similarity definitions for mixed data objects [1]. We divided the development process of such systems in 15 detailed steps where mandatory design choices exist. One of the most crucial steps thereof regards the *algorithmic mapping* of the 2D object arrangement to a similarity function. One approach for this algorithmic mapping is the calculation of attribute weights, depending on the correlation of attributes to the user-defined object arrangement, which can subsequently be utilized for the creation of similarity functions for mixed data.

*e-mail: juergen.bernard@igd.fraunhofer.de
†e-mail: david.sessler@igd.fraunhofer.de
‡e-mail: michael.behrisch@uni-konstanz.de
§e-mail: marco.hutter@igd.fraunhofer.de
¶e-mail: tobias.schreck@uni-konstanz.de
‖e-mail: joern.kohlhammer@igd.fraunhofer.de

IEEE Symposium on Visual Analytics Science and Technology 2014
November 9-14, Paris, France
978-1-4799-6227-3/14/$31.00 ©2014 IEEE

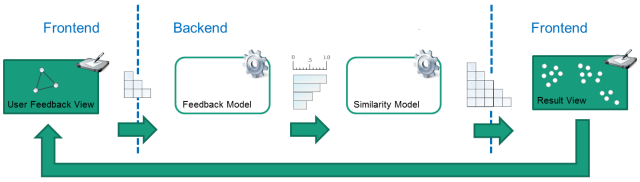

Figure 1: Data flow for visual-interactive similarity definition systems. Object arrangements are algorithmically mapped to distance matrices. Hence, the feedback model calculates attribute weightings which are utilized to learn similarity functions. A visual representation completes the sense-making loop.

In this work, we present a feedback model which generates weightings for categorical attributes based on user-defined feedback objects. For this purpose, we will showcase the results of empirical test cases which reveal challenges in the algorithmic mapping of categorical attributes (see Figure 3). In particular, we show that the cardinality of categorical attributes, and the number of feedback objects has an impact on the 'expressiveness' of the attribute weighting. Our feedback model overcomes these challenges by taking the results of the study into account. A repetition of the test cases shows a significant improvement of the algorithmic mapping.

2 A Feedback Model for Categorical Attributes

We present a feedback model for categorical attributes based on two stages. The first stage provides continuous weightings for categorical attributes based on user-defined object arrangements (see Section 2.1). In the second stage, we present an improved weighting (Section 2.3) which tackles the uncertainty aspect (Section 2.2).

2.1 Algorithmic Mapping of Object Arrangements

In the first stage, the object arrangement based on user feedback is mapped to attribute weights. Therefore, the system provides a distance matrix (DM) for every categorical attribute, calculated with the inverted Kronecker Delta function (see Figure 2 right). In return, users can define object arrangements (Figure 2 left). The feedback model maps the arrangement to pairwise Euclidean distances. Subsequently, the weight for each attribute is calculated by measuring the Pearson correlation between object DM and attribute DM.

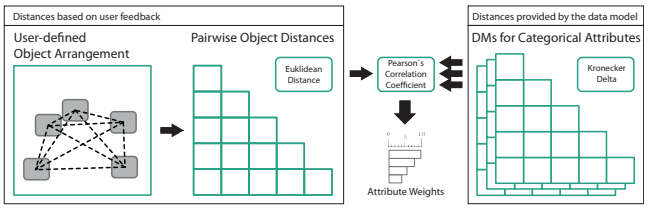

Figure 2: The feedback model maps the 2D arrangement of data objects to pairwise object distances. Attribute weights are calculated by correlating the distances with attribute distance matrices (DM).

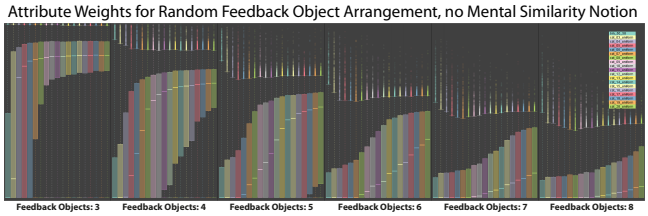

Figure 3: Boxplots showing the functional dependency between the number of feedback objects [3-8] (global x-axis), attribute cardinality [2-20] (nested x-axis), and the attribute weight generation [0.0-1.0] (y-axis). More feedback objects lead to less weighting uncertainty.

2.2 Empirical Test

Test Setup and Basic Assumption On the one hand, we assume an attribute to be part of the MSN if the (calculated) weighting is greater than zero. On the other hand whenever the attribute is not contained in the MSN, we can expect that it has a median weight of zero under generalisability assumption. We define an attribute as not being part of the MSN, if the feedback objects are arranged independently of the attribute properties. Therefore, in the test randomly picked objects are arranged randomly to ensure a) that no MSN is existing, and b) that the test covers every possible feedback arrangement. We use an artificial test data set generated such that cross correlations can be neglected. We chose the attribute weight as our independent variable. Our two dependent variables are the number of feedback objects (from 3 to 8) and the cardinality of the categorical attributes (from 2 to 20). To ensure generalisability, we repeat our tests 10.000 times for each experiment condition.

Test Results and Implications Figure 3 shows the uncertainty of categorical attribute weights represented with boxplots. Due to the multiple re-runs of the tests, we are able to use the median as a descriptive means for reasoning. We can derive two general insights. Firstly, the uncertainty of categorical attributes decreases with the number of feedback objects (global x-axis). Secondly, a larger cardinality of the attributes increases the uncertainty (nested x-axis). To make this point more clear, we see low weighting uncertainty if either the number of feedback objects is rather high (8 or higher), or if the attribute cardinality is rather low. More specifically, the median stays zero for an amount of feedback objects higher than 10 (not included in the Figure 3). However, although not being part of a MSN, for a clearly observable amount of experiment conditions *the median is higher than zero*.

2.3 Optimization of the Algorithmic Mapping

We aim to tackle the identified problem by optimizing the weight generation process. More specifically, we can take advantage of measured uncertainty medians as a minimum threshold t on the credibility of the attribute weight. If an attribute weight is lower than t, we have to assume that the weighting does not reflect the MSN precisely. On the other hand, if the attribute weighting is higher than t we can assume that the attribute is (to a certain degree) relevant for the MSN. Our approach eliminates the uncertainties by rescaling the attribute weightings according to the general understanding described above. Particularly, the new zero weighting value corresponds to the previously calculated t, while the maximum value is kept untouched. Weightings below t are set to zero. Since our approach relies on empirically measured values for t, we decided to store the values in a static look-up table, rather than calculating a (non-linear) regression model. Thus, we provide explicitly known values for t, for the finite number experiment conditions that reveal values above zero.

2.4 Improved Test Results

To measure the improvement of the optimization step, we created a test data set with identical parameters. In contrary to the latter test case, we included our improvement algorithm. The results can be seen in Figure 4. One can see that the newly calculated uncertainty medians in these boxplots can be found in very low value ranges

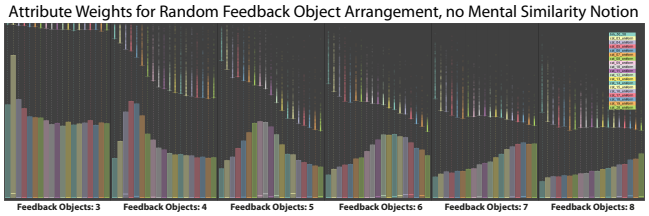

Figure 4: Repetition of the empirical test as described in Figure 3 with an incorporation of the algorithmic mapping optimization. The weighting uncertainty is reduced significantly for all experiment conditions (the number of feedback objects and the attribute cardinality).

(close to zero) for all experiment conditions. Hence, we could improve the overall performance significantly.

3 Discussion and Conclusion

While we have shown that the number of feedback objects and the cardinality of categorical attributes have a functional dependency with the degree of uncertainty, other dependent variables exist.

In the course of our research we also took the geometry of the object arrangement into account. While we still have to investigate whether or not participants will use geometrically regular arrangement aspects to reflect their MSN, we also conducted an empirical test based on regular object arrangements. Here, triangles for three, squares for four objects, etc. were used for the calculation of uncertainty weightings. Recap that again no MSN was assumed. The result can be seen on the attached poster. Due to the finite number of possible arrangement configurations, the earlier continuous weighting range gets discretized into a finite number of possible weights. But still the general insights of the earlier experiments (cf. Section 2.2) remain the same.

As another dependent variable, we are investigating the impact of non-uniformly distributed categorical attributes. The first results can be seen in Figure 5. We kept the number of feedback objects stable (4 objects) and tested attributes with cardinalities of 6, 9, and 15 (global x-axis). We tested the following Gaussian distributions: uniform, σ =3, 2, 1, 0.5, 0.3) (nested x-axis). Again, we can perceive higher uncertainties for higher cardinalities. Moreover, it can be seen that with a decreasing σ the uncertainty decreases. In other words: *diverse histograms have lower uncertainty*. One possible interpretation is that attributes with a high diversity tend to behave similarly to uniformly distributed attributes with a lower cardinality. Additionally, we can infer that our presented optimization step is robust with respect to non-uniform distributions down to a standard deviation of $\sigma = 2$.

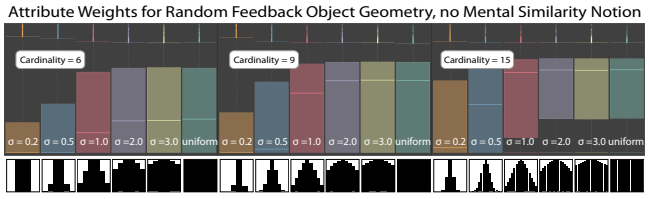

Figure 5: The object distribution within categorical attributes is another dependent variable. For the cardinalities 6, 9, and 15, Gaussian distributions based on different σ were generated and tested. The uncertainty decreases for standard deviations below $\sigma = 2$.

References

[1] J. Bernard, D. Sessler, T. Ruppert, J. Davey, A. Kuijper, and J. Kohlhammer. User-based visual-interactive similarity definition for mixed data objects-concept and first implementation. In *Int. Conf. in Central Europe on Comp. Graph., Vis. and Comp. Vis. (WSCG)*, 2014.

[2] J. Liu, E. T. Brown, R. Chang, and C. E. Brodley. Dis-function: Learning distance functions interactively. In *VAST*, pages 83–92. IEEE, 2012.

[3] G. M. H. Mamani, F. M. Fatore, L. G. Nonato, and F. V. Paulovich. User-driven feature space transformation. In *Proceedings of EuroVis*, pages 291–299. Eurographics Association, 2013.