EuroVis Workshop on Trustworthy Visualization (TrustVis) (2019)
R. Kosara, K. Lawonn, L. Linsen, and N. Smit (Editors)

Towards Supporting Interpretability of Clustering Results with Uncertainty Visualization

C. Kinkeldey, T. Korjakow and J. J. Benjamin

Human-Centered Computing Research Group, Freie Universität Berlin, Germany

Abstract

Interpretation of machine learning results is a major challenge for non-technical experts, with visualization being a common approach to support this process. For instance, interpretation of clustering results is usually based on scatterplots that provide information about cluster characteristics implicitly through the relative location of objects. However, the locations and distances tend to be distorted because of artifacts stemming from dimensionality reduction. This makes interpretation of clusters difficult and may lead to distrust in the system. Most existing approaches that counter this drawback explain the distances in the scatterplot (e.g., error visualization) to foster the interpretability of implicit information. Instead, we suggest explicit visualization of the uncertainty related to the information needed for interpretation, specifically the uncertain membership of each object to its cluster. In our approach, we place objects on a grid, and add a continuous "topography" in the background, expressing the distribution of uncertainty over all clusters. We motivate our approach from a use case in which we visualize research projects, clustered by topics extracted from scientific abstracts. We hypothesize that uncertainty visualization can increase trust in the system, which we specify as an emergent property of interaction with an interpretable system. We present a first prototype and outline possible procedures for evaluating if and how the uncertainty visualization approach affects interpretability and trust.

CCS Concepts

• ***Human-centered computing*** → ***Visualization design and evaluation methods;*** *Interface design prototyping; HCI theory, concepts and models;*

1. Introduction

Interpretability of Machine Learning (ML) models and outputs is crucial for understanding ML based systems, which affect both the mundane and geopolitical issues of our lives in various ways. The way we interact with such systems today typically does not support the interpretive work necessary to make sense of their output, especially for people with a non-technical background [Lip16]. A lot of the predictive power of ML comes from describing the world through high-dimensional features. Results derived from various ML algorithms and techniques often also exist in this space, and are therefore inherently hard to visualize. Dimensionality reduction techniques are used to reduce the data to 2D and make it usable in a visualization. Due to information loss, this procedure causes different kinds of artifacts in the data and makes interpretation of cluster results in scatter plots difficult [Aup07].

A way to support interpretability of algorithmic systems are *post-hoc techniques* [Lip16, Mil17]. They provide supplemental information to support interpretation of the outcome of computation, e.g., showing the top n elements of a cluster or saliency maps. However, techniques such as textual annotations, which are supposed to be beneficial for interpretation, are hard to apply in scatterplots where visual layout is unstructured. For this reason we suggest an approach that, instead of showing the original 2D space and trying to explain the distortion, positions the data points on a regular grid; which allows for more complex additional visual features. To inform about the uncertainty of each object to belong to its cluster, we compute its degree of membership to the cluster which can be interpreted as confidence in the clustering. We visualize a shaded relief to represent the membership distribution of the clusters, leading to a *cluster topography* (treating the term "topography" as equivalent to "relief"). Thereby, we seek not to supplement the visualization interface with separate elements (e.g., parameter selection), but instead attempt to enrich the visualization with information about uncertainty. We hypothesize that this approach increases trust in the outcome because it supports the interpretive work that is already required for our visualization, rather than introducing clutter in the form of overly technical features such as parameter selection that are also inherently difficult to interpret. We have developed this approach in a use case, where we perform topic extraction based on Latent Semantic Analysis (LSA) [KWR15] to uncover abstract topics in textual abstracts of research projects. In this context, we showcase a low-fidelity prototype that includes a cluster visualization with an integrated "topography" representing the distribution of uncertain object memberships within the cluster.

DOI: 10.2312/trvis.20191183

In the following we present our contributions to the workshop which are the proposal of an emergent notion of interpretability and trust for information visualization as well as the presentation of a prototype implementing a novel visualization approach for clustering results we name "cluster topography".

2. Interpretability and Trust in Machine Learning

Interpretability and trust can be seen as fundamentally intertwined for informed, self-driven use of ML systems by people with various backgrounds. However, multiple definitions exist for these concepts. Hoff and Bashir suggest that trust requires a diversified understanding. Based on a literature review of empirical research, they argue for distinguishing between dispositional, situational, and learned trust [HB15]. For this contribution, we consider the sublevel of *dynamic learned trust* (e.g., where an operator's preexisting knowledge and the automated system performance and design features meet). This level of trust is representative for the scenario of non-technical experts engaging with a machine-learning informed visualization. Additionally, Chuang et al. [CRMH12] have suggested guidelines for model-driven visualizations that are based on trust and interpretation as main criteria (as opposed to formal model quality measures). They define interpretation "[...] as the facility with which an analyst makes inferences about the underlying data" and trust as "[...] the actual and perceived accuracy of an analyst's inferences". Furthermore, there is evidence that awareness of underlying uncertainty can increase trust of people interacting with Visual Analytics tools [SSK*16].

Therefore, we do not consider dynamic learned trust as an outcome of a 'persuasive' interface, but rather as an emergent property of interaction, dependent on how people may meaningfully observe uncertainty and interact with an interpretable system. Regarding the latter, we base our approach to interpretability on the work of Lipton and Miller, who argue for a theory-based understanding of interpretability. Lipton states that while interpretability is frequently invoked in a "quasi-mathematical" manner, it is also ill-defined [Lip16]. Lipton distinguishes between the type of model transparency that an interpretable ML system may achieve, as well as the types of post-hoc techniques (e.g., visualization, text explanation, local explanation, explanation by example) that can make outcomes more interpretable. Miller adds to this observation that interpretability in ML research is commonly not based on theories of interpretation from the social sciences [Mil17]. As a consequence, Miller suggests that post-hoc interpretability techniques consider causal as well as contextual attribution. Therefore, we posit that an interpretable system supports dynamic learned trust if it (1) supports humans in making causal inferences about data processing and its effects (e.g., uncertainty) in a way that (2) reflects the context of human and system.

3. Visualization As A Post-Hoc Technique

Visualization is commonly used to communicate clustering results, typically in the form of a 2D scatterplot (Figure 1, left). This requires *dimensionality reduction (DR)* techniques such as *t-SNE* [MH08] to reduce the dimensionality of the data to 2D while preserving as much of its structure from the high dimensional attribute space as possible. However, because of the nature of DR, there are unavoidable artifacts that distort the distances between points in the reduced space [Aup07]. This means for instance that two pairs of data points in the same distance to each other are not necessarily equally similar. This intrinsic uncertainty can be misleading and erroneous, or result in overly confident interpretations; hence necessitating considerations of how to improve interpretability.

There are visual approaches that extend traditional representations of clustering results as scatterplots in order to counter the negative effects of DR. A common strategy is to represent artifacts caused by DR algorithms measures. They describe how the distances in the projected (2D) space relate to the distances in the n-dimensional space. Aupetit presents an interactive visualization to show distortions in the projected space with colored voronoi cells, termed as *proximity-based visualization* [Aup07]. Along the same lines, Heulot et al. modify this technique using interpolation instead of cells (*ProxiViz*) [HAF12]. The results of a controlled experiment [HFA17] suggest that for local tasks such as identification of outliers and clusters, people's responses were more accurate with ProxiViz than with a simple scatter plot. Martins et al. [MCMT14] follow a similar approach and visualize a set of quality metrics in the projected space, mainly referring to preservation of neighborhoods compared to the high-dimensional space. As part of the development of guidelines based on interpretability and trust, Chuang et al. [CRMH12] present the *Stanford Dissertation Browser* that displays academic departments by similarity of their topics. The authors address the problem of artifacts with a circular visualization that makes distortions in the projected space more transparent. They demonstrate how this can help detect missing plausibility in clustering results.

All in all, we can state that existing approaches mainly focus on making distortions in the projected space transparent, a strategy that was proven to be useful for low level tasks such as cluster identification and outlier detection. However, we hypothesize that this focus expresses a technical view on the data that may not be intuitive, especially for non-technical experts. For increasing trust not only in the result but in the machine learning system we see the need to explore new visual techniques fulfilling the requirements for post-hoc techniques as introduced above.

3.1. Motivating Use Case: Semantic Clustering of Research Projects Using Topic Extraction

This work is motivated by the development of a visualization tool for research projects. The people engaging with our visualization application are non-technical experts from a natural history research institution. The institution is nominally interdisciplinary, however, in practice knowledge is generated at workgroup level and is rarely shared across the disciplines that make up the various groups. Our application, therefore, is intended to showcase thematic overlaps between research projects, knowledge transfer activities and the use of infrastructures such as collections or labs. The goal is to encourage cross-disciplinary collaboration and sustainable use of generated knowledge. The most prominent feature for our application is a cluster visualization of research projects, as the latter are universally seen as the foundation for knowledge at the institution. The cluster visualization is the output of a topic

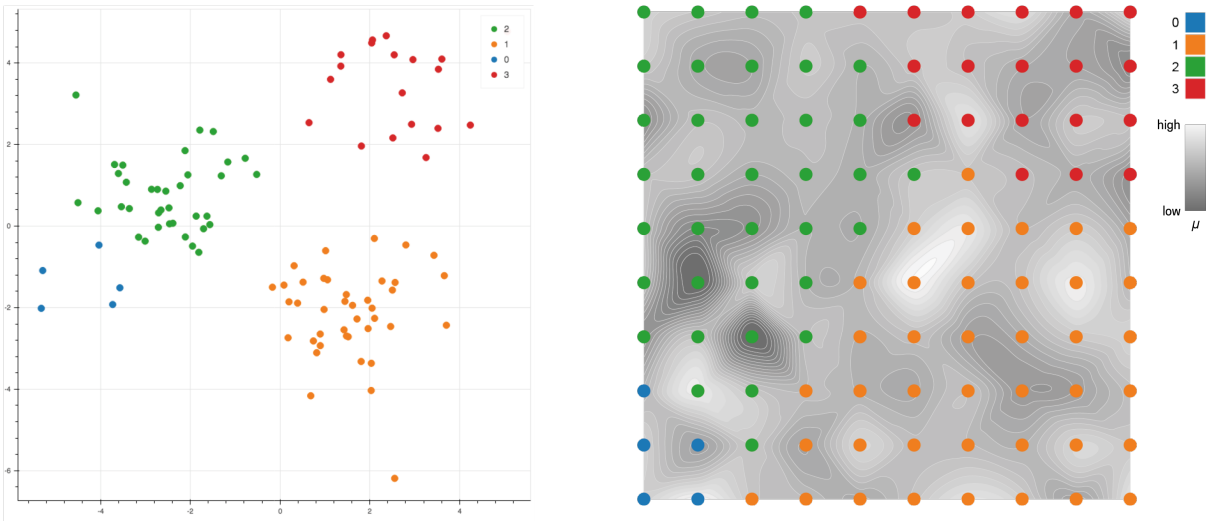

Figure 1: *Left: Scatterplot displaying 100 research projects clustered by topics extracted from scientific abstracts (see subsection 3.1) in four clusters (distinguished by color). Right: Our approach represents the same 100 projects on a grid, using the same colors for the clusters, and adds uncertain cluster membership μ as cluster "topography" (in grey scale, with white representing the highest degree of cluster membership = lowest uncertainty).*

modelling pipeline we have implemented for this purpose [†]. The data source for our application are publicly available descriptions of research projects funded by the central research funding organization in Germany [‡]. Our pipeline uses a preprocessing step, where project abstracts get vectorized using a term frequency-inverse document frequency weighting scheme (Tf-idf), followed by a Latent Semantic Analysis (LSA) to identify possible topics, k-means for clustering, and t-Stochastic Neighbor Embedding (t-SNE) or Linear Discriminant Analysis (LDA) for dimensionality reduction of the result to a 2D space. The goal of our cluster visualization is to allow for meaningful inferences on thematic overlaps between research projects; with dynamic learned trust as a property of the interaction with our visualization.

3.2. Visualization Approach: Cluster Topography

Our visualization approach is grounded in the general hypothesis that informing about the uncertainty in the data has the potential to increase dynamic learned trust in the application. Clusters are an abstract concept, intrinsically described by the objects forming the cluster and thus not straightforward to interpret. In our use case, one of the major goals was to inform about the uncertainty in the clustering results, i.e., about the uncertain membership of the objects in each cluster which can be interpreted as confidence in or reliability of the result. This information is not part of the actual end result of the topic extraction pipeline but can be obtained from an intermediary stage of the pipeline: we define an object's membership of its cluster as the closeness to the cluster centroid in the n-dimensional topic space (computed as the Euclidean distance between the object and the cluster centroid). Since the distances in the reduced 2D space have limited expressiveness (as discussed above) we decided to place the objects on a grid using a Jonker-Volgenant linear optimization algorithm [JV87] that shifts each point to its closest grid point while aiming to preserve the topological characteristics of the unstructured point set. In doing so, we lose the information of the objects' exact position but avoid clutter (a common drawback of scatter plots) which makes more complex visualizations possible. We see the use of metaphors to represent cluster attributes as key to enhance interpretability of clusters. That is why the distribution of cluster membership μ of the objects within each cluster is visualized in the form of a shaded surface that we interpolate between the objects evoking the notion of a land surface (Figure 1). For instance, objects on top of a "hill" are best represented by their cluster whereas objects in a "basin" do not "fit" well into their cluster. We hypothesize that the use of this metaphor increases the intuitiveness of the visualization of cluster memberships.

In addition to the cluster topography, we provide the five most frequent topics for each object as tooltips. In interpretive terms those can be seen as the topics that are most responsible for the object being assigned to its cluster. Browsing the lists of influential topics can provide information about why objects are in the same cluster, which can potentially strengthen the dynamic learned trust in the result. In the example shown in Figure 2 the observation that the highlighted research projects are located in the "basin" of the green cluster leads to the question if they do not "fit" into the cluster. Hovering over the two points reveals that they do not share a main topic with the cluster but share most of the topics between them. Based on the knowledge about the compatibility of the topics a person can now decide if the low membership is a false alarm (if the topics are related although the algorithm claims otherwise)

† https://github.com/FUB-HCC/IKON-backend
‡ https://gepris.dfg.de

or if the objects are actual outliers. As discussed above, this dimension of trust assumes a pre-existing knowledge of the operator. Therefore, displaying topic words in combination with the cluster topography supports the interpretability of how machine learning performance and expert knowledge relate and can potentially raise the trust in the topic pipeline used here.

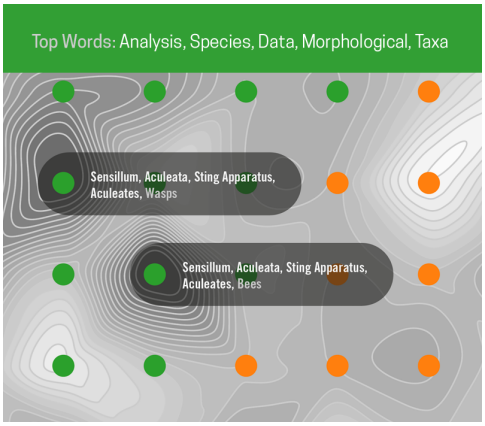

Figure 2: *Detail view of the cluster topography with two research projects in a "basin" of the cluster's topography. Their main topics (grey labels) are almost identical but do not overlap with those of the cluster (list in the green area on top).*

4. Discussion

As discussed above, visualization of high-dimensional spaces as used for topic clustering requires a reduction in dimensionality. Thereby, artifacts are introduced which inevitably distort the interpretive context. We posit that textual explanations in form of annotations can also be seen as artifacts in that sense, injecting additional information into the interpretive context. This suggests a fundamental characteristic of interpretability techniques: their use is generative of artifacts which become integrated into whatever is sought to be made more interpretable. In this sense, each specific interpretability technique has a "trade-off" which may affect learned trust.

In comparison to approaches that display distortions caused by the projection from high- to low-dimensional space [HFA17], we hypothesize that our approach will prove to be more intuitive related to interpretation. We do not try to make the relative locations of objects in the projected space interpretable, instead, we draw from information from an earlier stage in the pipeline and explicitly represent it in a metaphorical way. Our approach can be seen as addition to a standard scatterplot (potentially as linked views, in which hovering reveals the location of a point in the other view) which may be a way to combine the advantages from both visualizations.

A clear limitation of the approach in its current form is that, based on the information about the uncertainty of the objects' cluster memberships, pairwise comparisons between the objects cannot be made. Since we chose the Euclidean distance in the topic space as a metric of similarity, there can be an infinite number of points on an n-dimensional sphere around the cluster center which exhibit the same uncertainty, but may be characterized by substantially different topics. So far we have not explored the question if people assume a direct semantic connection between objects on the same "height" of the relief, but this could potentially be misleading. A part of future work will be to look into alternative cluster membership measures such as the silhouette score [§]. Related to this, another critical aspect is that we use cubic interpolation to create the relief. This leads to smooth isolines, but is not grounded in assumptions how the space may look like between the data points. This is certainly an aspect that needs further attention.

5. Conclusion

We presented a visualization approach that we suggest as a post-hoc technique to enhance interpretability of clustering results. The motivation for this work lies in the basic level of research on the evaluation of interpretability of machine learning systems; that is, existing work does not go beyond the accuracy of cluster and outlier detection in the data. We see the need to go further and to develop methods that assess actual interpretation of the content as well as people's trust in the system they use. In general, we suggest to explicitly visualize the uncertainty related to the information needed for interpretation of the results. In our case, we visualized uncertain memberships of research projects to their clusters as a continuous cluster "topography".

The discussion on interpretability and trust in machine learning is currently lacking concrete use cases for non-technical experts. Therefore, we believe that an evaluation of our approach will contribute to the further development of this field. As we are focusing on the effects of our post-hoc techniques on dynamic learned trust, qualitative evaluation methods should be pursued. Currently, we are considering formative evaluation methods in the form of co-discovery user tests with varying versions of our low-fidelity prototypes. In this way, researchers may discuss similarities between research projects, and actively probe whether our cluster topography is supportive. Their interactions with each other and the system, and how prior knowledge and system features correlate, can then be analyzed for the kind of interpretive work that our application supports. Possible avenues for summative evaluation are longitudinal field studies, in which we observe researchers interacting with the cluster visualization over a sustained amount of time and log their interactions and conduct a series of interviews.

At the workshop we intend to discuss the following issues:

- Visualization as post-hoc interpretability technique
- Potential and limitations of the cluster topography approach
- Evaluation methods for capturing interpretability and trust

6. Acknowledgements

This work is supported by the German Federal Ministry of Education and Research, (BMBF), grant 03IO1633 ("IKON – Wissenstransferkonzept für Forschungsinhalte, -methoden und -kompetenzen in Forschungsmuseen").

§ https://en.wikipedia.org/wiki/Silhouette_(clustering)

References

[Aup07] AUPETIT M.: Visualizing distortions and recovering topology in continuous projection techniques. *Neurocomputing 70*, 7-9 (2007), 1304–1330. 1, 2

[CRMH12] CHUANG J., RAMAGE D., MANNING C., HEER J.: Interpretation and trust: designing model-driven visualizations for text analysis. In *Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems - CHI '12* (Austin, Texas, USA, 2012), ACM Press, p. 443. URL: http://dl.acm.org/citation.cfm?doid=2207676.2207738, doi:10.1145/2207676.2207738. 2

[HAF12] HEULOT N., AUPETIT M., FEKETE J.-D.: Proxiviz: an interactive visualization technique to overcome multidimensional scaling artifacts. *Proceedings of IEEE InfoVis, poster* (2012). 2

[HB15] HOFF K. A., BASHIR M.: Trust in Automation: Integrating Empirical Evidence on Factors That Influence Trust. *Human Factors 57*, 3 (May 2015), 407–434. URL: https://doi.org/10.1177/0018720814547570, doi:10.1177/0018720814547570. 2

[HFA17] HEULOT N., FEKETE J.-D., AUPETIT M.: Visualizing dimensionality reduction artifacts: An evaluation. *arXiv preprint arXiv:1705.05283* (2017). 2, 4

[JV87] JONKER R., VOLGENANT A.: A shortest augmenting path algorithm for dense and sparse linear assignment problems. *Computing 38*, 4 (1987), 325–340. 3

[KWR15] KARL A., WISNOWSKI J., RUSHING W. H.: A practical guide to text mining with topic extraction. *Wiley Interdisciplinary Reviews: Computational Statistics 7*, 5 (2015), 326–340. 1

[Lip16] LIPTON Z. C.: The Mythos of Model Interpretability. *arXiv:1606.03490 [cs, stat]* (June 2016). arXiv: 1606.03490. URL: http://arxiv.org/abs/1606.03490. 1, 2

[MCMT14] MARTINS R. M., COIMBRA D. B., MINGHIM R., TELEA A. C.: Visual analysis of dimensionality reduction quality for parameterized projections. *Computers & Graphics 41* (2014), 26–42. 2

[MH08] MAATEN L. V. D., HINTON G.: Visualizing data using t-sne. *Journal of machine learning research 9*, Nov (2008), 2579–2605. 2

[Mil17] MILLER T.: Explanation in Artificial Intelligence: Insights from the Social Sciences. *arXiv:1706.07269 [cs]* (June 2017). arXiv: 1706.07269. URL: http://arxiv.org/abs/1706.07269. 1, 2

[SSK*16] SACHA D., SENARATNE H., KWON B. C., ELLIS G., KEIM D. A.: The role of uncertainty, awareness, and trust in visual analytics. *IEEE transactions on visualization and computer graphics 22*, 1 (2016), 240–249. 2