Eurographics Conference on Visualization (EuroVis) (2013)
M. Hlawitschka and T. Weinkauf (Editors)

Short Papers

Information Retrieval Perspective to Interactive Data Visualization

J. Peltonen[1], M. Sandholm[1], and S. Kaski[1,2]

[1]Helsinki Institute for Information Technology HIIT, Department of Information and Computer Science, Aalto University, Finland
[2]Helsinki Institute for Information Technology HIIT, Department of Computer Science, University of Helsinki, Finland
Email: {jaakko.peltonen, samuel.kaski}@aalto.fi, msandhol@gmail.com

Abstract

Dimensionality reduction for data visualization has recently been formulated as an information retrieval task with a well-defined objective function. The formulation was based on preserving similarity relationships defined by a metric in the input space, and explicitly revealed the need for a tradeoff between avoiding false neighbors and missing neighbors on the low-dimensional display. In the harder case when the metric is not known, the similarity relationships need to come from the user. We formulate interactive visualization as information retrieval under uncertainty about the true similarities, which depend on the user's tacit knowledge and interests in the data. During the interaction the user points out misses and false positives on the display; based on the feedback the metric is gradually learned and the display converges to visualizing similarity relationships that correspond to the tacit knowledge of the user.

Categories and Subject Descriptors (according to ACM CCS): H.5.m [Information interfaces and presentation (e.g., HCI)]: Miscellaneous

1. Introduction

We study interactive data exploration with scatter plots. Traditional plots of feature pairs only reveal part of the structure in multivariate data; to show high-dimensional data on a scatter plot, *nonlinear dimensionality reduction* (NLDR) is often applied. What is relevant in data is normally *not known a priori*; it depends on the user's interest. Thus simple visualization using naive assumptions about what data properties are important will not work well; *interactive visualization* should be used to let the user give feedback about what is relevant. We introduce a novel information retrieval based approach to interactive data exploration with scatter plots.

A recent method [VK07b, VPN*10, PK11] formalizes the case where the user is interested in neighborhood relationships between data points. A static (non-interactive) visualization with scatter plots is formalized as a rigorous *information retrieval task* where the user retrieves neighborhood relationships based on the display; the display is optimized to minimize errors between retrieved neighbors and true neighborhoods in the input space. The optimized display is then a faithful representation of the data in the well-defined sense of yielding few errors in the visual information retrieval. The formalism yields the Neighbor Retrieval Visualizer (NeRV) method which has outperformed several methods [VPN*10]; the methods Stochastic Neighbor Embedding (SNE; [HR02]) and t-Distributed SNE [vdMH08] can also be interpreted as special cases of the formalism.

A general way to encode what aspects of data are relevant to a user is to define the *metric between data* so that differences between data depend on the relevant aspects. Most NLDR methods require known distances or a known metric; NeRV and related methods use the known metric to compute input space neighborhoods between high-dimensional data. When the metric is not known a priori, a natural approach is to learn it by interacting with the user. We assume the user's interaction is based on an underlying metric that encodes the user's tacit knowledge and interests in the data; we call input neighborhoods in this tacit metric *true neighborhoods*. The task of the visualization system then is to learn the metric and to compress the data to the display. We give a rigorous solution: **we introduce an interactive visualization method optimized to serve the user in the rigorously defined task of retrieving true neighbors from the scatter plot**. We infer the metric iteratively from feedback in the

DOI: 10.2312/PE.EuroVisShort.EuroVisShort2013.049-053

user's retrieval task, and optimize the arrangement of data on the display for each iteration, to serve information retrieval in the inferred metric. We concentrate on pairwise similarity and dissimilarity feedback. We introduce a formulation of interactive visualization as *information retrieval under uncertainty of the user preferences*. Our method can be seen as the full interactive extension of the NeRV formalism.

Earlier methods for interactive data exploration with scatter plots include, for example, the Grand Tour [Asi85], and methods where the user explicitly decides what dimensions to plot. Recently, systems that try to learn from observation-level interactions how the user wishes to arrange data [EHM*11, EFN12, BLBC12] have been proposed. An advantage of our system is that the whole interactive process is optimized for the rigorous user task of neighbor retrieval.

2. Interactive visualization as information retrieval

Scatter plot visualization of multivariate data is often done by applying NLDR; several methods exist [TdSL00, BN02, RS00]. Many NLDR methods do not perform well in visualization [VK07a]; they have not been designed to reduce dimensionality beyond the effective dimensionality of the data manifold, and are not good at compressing data onto a low-dimensional display. We review a formalization of NLDR which has proven successful in static visualization [VPN*10], and we then extend it to interactive visualization.

Let $\{\mathbf{x}_i\}_{i=1}^N$ be a set of input data samples. Let each sample i have an unobserved *true neighborhood* p_i, which is a distribution telling for each neighbor j the probability $p_{j|i}$ that j is chosen as a neighbor to i. The user's true neighborhoods will be learned from feedback. The goal is to create output coordinates $\{\mathbf{y}_i\}_{i=1}^N$ for the data suitable for visual neighbor retrieval. On the display an *output neighborhood* q_i can be defined around each sample as probabilities $q_{j|i}$, in this paper $q_{j|i} = \frac{\exp(-||\mathbf{y}_i-\mathbf{y}_j||^2/\sigma_i^2)}{\sum_{k\neq i}\exp(-||\mathbf{y}_i-\mathbf{y}_k||^2/\sigma_i^2)}$, where $||\cdot||^2$ is squared Euclidean distance on the display; $q_{j|i}$ is the probability that an analyst starting from a central point i will consider j similar on the display. This simple mathematical form can be replaced by more advanced user models if available.

All properties of high-dimensional data cannot be represented on a low-dimensional scatter plot. Two kinds of errors will happen (Fig. 1, top): *misses* are true neighbors of a point i (high $p_{j|i}$) that are not neighbors on the display (low $q_{j|i}$). *False neighbors* are neighbors on the display (high $q_{j|i}$) that are not true neighbors (low $p_{j|i}$). Misses and false neighbors can have a different cost to the analyst. The display should be optimized to minimize the total cost of errors. It has been shown [VPN*10] that the total cost of misses corresponds to the information retrieval measure *recall*, and the total cost of false neighbors corresponds to *precision*. The measures have been generalized to divergences between probabilistic neighborhoods [VPN*10]: the Kullback-Leibler divergence $D(p_i,q_i) = \sum_{j\neq i} p_{j|i} \log \frac{p_{j|i}}{q_{j|i}}$ is a generalization of recall and $D(q_i,p_i) = \sum_{j\neq i} q_{j|i} \log \frac{q_{j|i}}{p_{j|i}}$ is a generalization of precision. The total information retrieval cost C_{NeRV} of misses and false neighbors is then

$$C_{\mathrm{NeRV}} = \lambda \mathbb{E}_i[D(p_i,q_i)] + (1-\lambda)\mathbb{E}_i[D(q_i,p_i)] \quad (1)$$

where $\mathbb{E}_i$ denotes expectation over the query points i. The parameter λ controls the precision-recall tradeoff desired by the analyst: whether misses or false neighbors are more important to avoid. Different tradeoffs yield different optimal displays (Fig. 1, bottom). In our interactive sessions we emphasize precision (λ near 0) since then intermediate plots are locally well arranged which can make it easier to browse data as the analyst is not distracted by false neighbors. To optimize a visualization, (1) must be optimized with respect to the output coordinates $\mathbf{y}_i$ that define the neighborhoods $q_i = \{q_{j|i}\}$. The previous static visualization approach [VPN*10] can do this only if the true neighborhoods $p_i = \{p_{j|i}\}$ for each data point i are known. We treat the harder case when the true neighborhoods are unknown; we now extend the approach to unknown true neighborhoods.

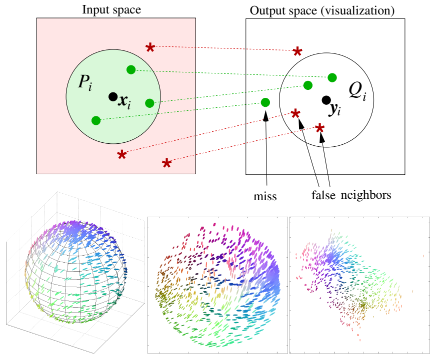

Figure 1: ***Top:*** *errors in visual information retrieval for query point i. P_i denotes points with high true neighborhood probability, Q_i denotes points with high neighborhood probability on the display.* Misses *are true neighbors that are not neighbors on the display;* false neighbors *are neighbors on the display that are not true neighbors.* ***Bottom:*** *different tradeoffs between recall and precision (misses and false neighbors) yield different optimal 2D displays. Original 3D data (bottom left) are on a sphere surface; flattening the sphere (bottom center) avoids misses but yields false neighbors from opposite sides, cutting the sphere open (bottom right) avoids false neighbors but yields misses over the cuts.*

2.1. Interactive visualization optimized for information retrieval under uncertainty

Equation (1) can be computed only if the true neighborhoods $p_i = \{p_{j|i}\}$ are known. When the true neighborhoods are *unknown*, but evidence of them is available in the form of user feedback, the rigorous approach is to treat the true neighborhoods as missing values, and optimize the expectation of the cost function over the missing values. That is, we *optimize the visualization for information retrieval under uncertainty of the user's preferred similarities*. This is written as

$$\mathbb{E}[C_{\mathrm{NeRV}}] = \mathbb{E}_{\{p_i\}|F}\left[\lambda \mathbb{E}_i[D(p_i,q_i)] + (1-\lambda)\mathbb{E}_i[D(q_i,p_i)]\right] \tag{2}$$

where $\mathbb{E}_{\{p_i\}|F}$ denotes expectation over the possibilities for different true neighborhood distributions $\{p_i\}$; the expectation is over a posterior distribution of the possible neighborhood distributions, given the evidence from feedback F.

Assume the true neighborhoods $p_i = \{p_{j|i}\}$ have a simple form: they are a function of distances in an *unknown metric* of the original multivariate feature space, so that

$$p_{j|i} = \frac{\exp(-||\mathbf{x}_i - \mathbf{x}_j||^2_{\mathbf{A}}/\sigma_i^2)}{\sum_{k \neq i} \exp(-||\mathbf{x}_i - \mathbf{x}_k||^2_{\mathbf{A}}/\sigma_i^2)} \tag{3}$$

where $||\mathbf{x}_i - \mathbf{x}_j||^2_{\mathbf{A}} = (\mathbf{x}_i - \mathbf{x}_j)^\top \mathbf{A}(\mathbf{x}_i - \mathbf{x}_j)$ and $\mathbf{A}$ is the metric matrix. The expectation over possible true neighborhoods in (2) becomes an expectation over possible true metrics:

$$\mathbb{E}[C_{\mathrm{NeRV}}] = \mathbb{E}_{\mathbf{A}|F}\left[\lambda \mathbb{E}_i[D(p_i,q_i)] + (1-\lambda)\mathbb{E}_i[D(q_i,p_i)]\right] \tag{4}$$

where the true neighborhoods $p_i = \{p_{j|i}\}$ are now functions of the true metric which we denote by its associated matrix $\mathbf{A}$, and $\mathbb{E}_{\mathbf{A}|F}$ denotes expectation over a posterior distribution of metrics $\mathbf{A}$ given the feedback F.

Equation (4) is an expectation (integral) of C_{NeRV} over the posterior distribution of metrics. We use a fast estimate of the integral. We infer a *variational approximation* $\hat{p}(\mathbf{A}|F)$ to the posterior $p(\mathbf{A}|F)$ as described in Section 2.2. We then approximate the integral by the value of C_{NeRV} at the mean $\mathbf{A}^* = \mathbb{E}_{\hat{p}(\mathbf{A}|F)}[\mathbf{A}]$ of the variational posterior. The variational distribution is unimodal and is optimized to contain much of the posterior mass, so the value of C_{NeRV} at $\mathbf{A}^*$ is a reasonable quick-and-dirty approximation to the integral. We write

$$\mathbb{E}[C_{\mathrm{NeRV}}] \approx \left[\lambda \mathbb{E}_i[D(p_i,q_i)] + (1-\lambda)\mathbb{E}_i[D(q_i,p_i)]\right]_{\mathbf{A}=\mathbf{A}^*} \tag{5}$$

where the true neighborhoods p_i are computed by (3) using the mean posterior metric $\mathbf{A}^*$ and the output neighborhoods q_i are computed from display coordinates $\{\mathbf{y}_i\}$ of the data as defined in Section 2. Equation (5) **measures the performance of a visualization in the information retrieval task of retrieving the true neighbors, corresponding to the analyst's tacit knowledge, from the display**. Equation (5) can be used as an *optimization criterion*, since it is a well-defined function of the display coordinates of the data.

Interactive optimization of the cost (5) performs the following three steps at each iteration. **1.** Infer the approximate posterior mode $\mathbf{A}^*$ of the metric from feedback received so far. **2.** Optimize the visualization for the neighborhoods p_i yielded by the metric $\mathbf{A}^*$. **3.** Show the new visualization and gather feedback from the analyst. The visualization is optimized by minimizing (5) with respect to the low-dimensional output coordinates $\mathbf{y}_i$ of each point i; we optimize them by conjugate gradient descent. The approach has a rigorous interpretation: the display is optimized for minimal expected cost of misses and false neighbors. We call the resulting interactive visualization method the *Interactive Neighbor Retrieval Visualizer* (Interactive NeRV).

2.2. Inference of the metric from feedback

We assume the analyst gives feedback on pairs of points, labeling them similar or dissimilar. We use a Bayesian approach to learn the metric from feedback. The amount of feedback needed depends on dimensionality, not on the number of visualized data. The metric is parameterized as $\mathbf{A} = \sum_{d=1}^{D} \gamma_d \mathbf{v}_d \mathbf{v}_d^\top$ where the $\mathbf{v}_d$ are basis vectors for the data and γ_d are weighting parameters that differentiate the possible metrics. In experiments we use the original basis of the data, therefore we learn weighted Euclidean metrics, which makes analysis of the results easy. Inferring $\hat{p}(\mathbf{A}|F)$ from a set of feedback pairs $F = \{(i,j,f_{ij})\}$ is then done by inferring the variational posterior approximation for the γ_d. The likelihood of a single feedback pair is defined as $p(f_{ij}|\mathbf{x}_i,\mathbf{x}_j,\mathbf{A},\mu) = (1+\exp(f_{ij}(||\mathbf{x}_i - \mathbf{x}_j||^2_{\mathbf{A}} - \mu)))^{-1}$, where μ is a threshold parameter and $f_{ij} = 1$ for a similar pair and -1 for a dissimilar pair. Given a Gaussian prior for the weighting parameters γ_d, and the likelihood terms for all feedback pairs as above, we can infer a variational approximation for the posterior of γ_d. Details of the update equations are omitted for brevity. Equivalent Bayesian updates have been used for metric learning without a visualization context [YJS07]; our novel contribution for the metric learning is to integrate the updates as a part of interactive optimization of the information retrieval cost $\mathbb{E}[C_{NeRV}]$.

3. Experiments

We evaluate our method in three ways: **1.** we evaluate the benefit of utilizing a visualization in finding good feedback pairs, **2.** we test whether the iterative interaction and metric learning help the user in the task of visual retrieval of relevant neighbors, and **3.** we present a small case study with a real user. In experiments 1 & 2, in each iteration 3 pairs of feedback are produced by an artificial mechanism: we compare the current visualization to known true neighborhoods and give the worst misses or false neighbors as feedback. We use three data sets in the experiments: articles published by researchers of a local research institute, a subset of the DARPA TIMIT phoneme data, and Wine from UCI machine learning repository. Each data set has additional noise features, assumed not to be beneficial for retrieving the true neighborhood relationships corresponding to user interests. We built a simple implementation of our approach, where the user is shown a scatter plot and he interacts by picking neighbors and non-neighbors, and can inspect data items by hovering over them with the mouse. Exhaustive search is not needed, only that the user finds some point pairs for feedback; we color previous feedback points to help. Our approach can be integrated in larger systems and combined with inspection tools like lenses, linked displays, glyphs etc., and with tools to annotate points or regions.

Figure 2 (top) shows, using an oracle user who always picks the worst miss or false neighbor (with respect to a known true neighborhood), that giving feedback based on the visualization improves metric learning compared to picking the pair randomly. Figure 2 (bottom) shows that visualization quality improves as the worst misses and false neighbors are pointed out, and that our information retrieval-based visualizer outperforms traditional multidimensional scaling (MDS) coupled to metric learning on two data sets. We measure for each visualization the area under the precision-recall curve; in the curve, the true neighbors are the 20 closest neighbors in the ground truth metric, the number of neighbors retrieved from the visualization is varied from 1 to 100, and mean precision and recall are computed at each number of retrieved neighbors. This experiment can be seen as transductive learning: by giving feedback to only a small amount of pairs, the overall accuracy of neighborhoods improves.

Figure 3 shows a small-scale user study using scientific articles as the data set. The user's goal was to arrange the scatter plot so that articles the user considers similar are close to each other on screen, by giving pairwise feedback. To help the user browse the points, we showed the title, year, and authors of the paper in a pop-up display when the user hovered over the corresponding point with the mouse. Points also changed color after the user gave feedback on them. The user gave 23 feedback pairs in total over 20 iterations. Figure 3 shows that as feedback was given, the metric improved and articles became arranged according to research fields.

4. Conclusions

We introduced an interactive visualizer that serves the user in an information retrieval task, finding neighborhood relationships corresponding to the user's tacit knowledge. The true neighborhoods are assumed to be expressible as a metric of the high-dimensional data. The user points out misses and false positives on the display, the metric is inferred from the feedback, and the display is optimized for information retrieval in the inferred metric; the display then iteratively converges to show similarity relationships relevant to the user. The whole system is rigorously quantifiable and optimizable by performance in the retrieval task. In experiments the visualizer learns metrics better than a non-visual mechanism, and shows relevant neighbors better than an alternative multidimensional scaling method coupled to the metric.

5. Acknowledgments

The work was supported by Academy of Finland, decisions 251170 (Finnish CoE in Computational Inference Research COIN), 252845 and 255725. Authors belong to COIN. Document data are derived from the following indices: Science Citation Index Expanded, Social Science Citation Index and Arts & Humanities Citation Index, prepared by Thomson Reuters®, Philadelphia, Pennsylvania, USA, ©Copyright Thomson Reuters®, 2011, and the Digital Library of the Association of Computing Machinery (ACM).

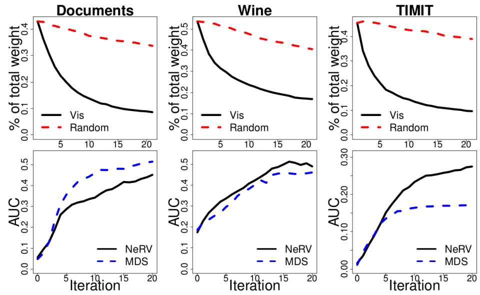

Figure 2: ***Top:*** *Feedback pairs chosen based on the visualization ("Vis") improve the metric learning compared to selecting the pair randomly in all data sets. The data sets include added irrelevant features containing only noise; we measure the portion of weight the metric assigns to the noise features. Both mechanisms decrease the importance of noise features, our method "Vis" does so faster.* ***Bottom:*** *Area under the precision-recall curves using the ground truth metric. Retrieval performance improves when we give feedback about errors of previous visualizations. Interactive NeRV outperforms the MDS based system on two data sets.*

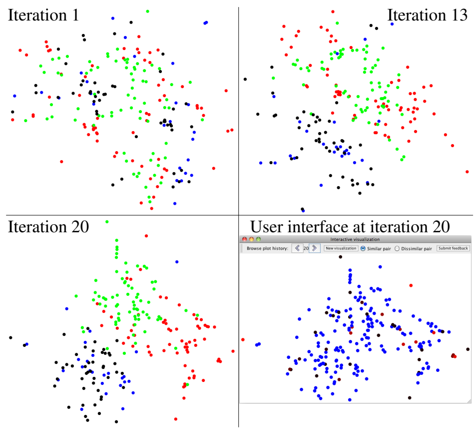

Figure 3: *Example visualizations from the user study. Colors show research fields: black=human-computer interaction, green=machine learning, red=complexity theory, blue=social psychology. Starting from the first visualization (top left), points become more arranged according to the hidden colors as we give more feedback (top right, bottom left). The interface (bottom right) shows data without labels, and feedback point pairs in red (brighter=more recent).*

References

[Asi85] ASIMOV D.: The grand tour: A tool for viewing multidimensional data. *SIAM Journal on Scientific and Statistical Computing 6*, 1 (1985), 128–143. 2

[BLBC12] BROWN E. T., LIU J., BRODLEY C., CHANG R.: Dis-function: Learning distance functions interactively. In *IEEE Conference on Visual Analytics Science and Technology (VAST)* (2012), IEEE, pp. 83–92. 2

[BN02] BELKIN M., NIYOGI P.: Laplacian eigenmaps and spectral techniques for embedding and clustering. In *Advances in Neural Information Processing Systems 14* (Cambridge, MA, 2002), Dietterich T. G., Becker S., Ghahramani Z., (Eds.), MIT Press, pp. 585–591. 2

[EFN12] ENDERT A., FIAUX P., NORTH C.: Semantic interaction for visual text analytics. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (2012), CHI '12, ACM, pp. 473–482. 2

[EHM*11] ENDERT A., HAN C., MAITI D., HOUSE L., LEMAN S., NORTH C.: Observation-level interaction with statistical models for visual analytics. In *IEEE Conference on Visual Analytics Science and Technology (VAST)* (2011), IEEE, pp. 121–130. 2

[HR02] HINTON G., ROWEIS S. T.: Stochastic neighbor embedding. In *Advances in Neural Information Processing Systems 14*, Dietterich T., Becker S., Ghahramani Z., (Eds.). MIT Press, Cambridge, MA, 2002, pp. 833–840. 1

[PK11] PELTONEN J., KASKI S.: Generative modeling for maximizing precision and recall in information visualization. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics* (2011), Gordon G., Dunson D., Dudik M., (Eds.), vol. 15 of *JMLR W&CP*, JMLR, pp. 597–587. 1

[RS00] ROWEIS S. T., SAUL L. K.: Nonlinear dimensionality reduction by locally linear embedding. *Science 290* (2000), 2323–2326. 2

[TdSL00] TENENBAUM J. B., DE SILVA V., LANGFORD J. C.: A global geometric framework for nonlinear dimensionality reduction. *Science 290* (2000), 2319–2323. 2

[vdMH08] VAN DER MAATEN L., HINTON G.: Visualizing data using t-SNE. *Journal of Machine Learning Research 9* (2008), 2579–2605. 1

[VK07a] VENNA J., KASKI S.: Comparison of visualization methods for an atlas of gene expression data sets. *Information Visualization 6* (2007), 139–54. 2

[VK07b] VENNA J., KASKI S.: Nonlinear dimensionality reduction as information retrieval. In *Proceedings of AISTATS*07, the 11th International Conference on Artificial Intelligence and Statistics (JMLR Workshop and Conference Proceedings Volume 2)* (2007), Meila M., Shen X., (Eds.), pp. 572–579. 1

[VPN*10] VENNA J., PELTONEN J., NYBO K., AIDOS H., KASKI S.: Information retrieval perspective to nonlinear dimensionality reduction for data visualization. *Journal of Machine Learning Research 11* (2010), 451–490. 1, 2

[YJS07] YANG L., JIN R., SUKTHANKAR R.: Bayesian active distance metric learning. In *Proceedings of UAI-07, the Twenty-Third Annual Conference on Uncertainty in Artificial Intelligence* (2007), Parr R., van der Gaag L., (Eds.), AUAI Press, pp. 442–449. 3