EuroVis Workshop on Visual Analytics (2017)
M. Sedlmair and C. Tominski (Editors)

Combining Cluster and Outlier Analysis with Visual Analytics

Jürgen Bernard[1], Eduard Dobermann[1], Michael Sedlmair[2], and D. W. Fellner[1,3,4]

[1]TU Darmstadt, Darmstadt, Germany
[2]Universität Wien, Vienna, Austria
[3]Fraunhofer IGD, Darmstadt, Germany
[4]Graz University of Technology, Institute of Computer Graphics and Knowledge Visualization, Austria

Abstract

Cluster and outlier analysis are two important tasks. Due to their nature these tasks seem to be opposed to each other, i.e., data objects either belong to a cluster structure or a sparsely populated outlier region. In this work, we present a visual analytics tool that allows the combined analysis of clusters and outliers. Users can add multiple clustering and outlier analysis algorithms, compare results visually, and combine the algorithms' results. The usefulness of the combined analysis is demonstrated using the example of labeling unknown data sets. The usage scenario also shows that identified clusters and outliers can share joint areas of the data space.

CCS Concepts

•***Information systems*** → ***Data mining;*** •***Human-centered computing*** → ***Visual analytics;*** *Information visualization;* •***Theory of computation*** → *Active learning;* •***Computing methodologies*** → *Machine learning;*

1. Introduction

The identification of clustered objects is an important task in exploratory data analysis. The same holds for the identification of outlier objects, which is, to a certain degree, a counterpart to cluster analysis. Clustering is often used to aggregate the data, i.e., to replace the entire data set by information about a small set of representatives, e.g., to address problems in the era of big data (see, e.g., Jain's survey on clustering [Jai10]). In general, cluster analysis can be used as a means to make informed decisions about 'a series of generalizable' objects. In the terminology of the visual analytics community the granularity of analysis is raised from an elementary to a synoptic level [AA06]. In turn, outlier analysis is often used to explore unexpected or anomalous observations hidden in complex data sets [CBK09]. As such, this elementary analysis task is important to identify objects where generalizable statements about the data set hardly meaningful, or do not even exist.

Based on their definitions, clustering and outlier analysis support opposed analysis tasks. The interpretation of clustered objects and outliers seem to refer to disjoint areas of the data space. Accordingly, only few visual approaches exist where clustering and outlier analysis are used in combination. One research question that arises is (1) whether data objects identified by clustering algorithms and objects identified by outlier algorithms are really disjunct in any case? Another question is (2) whether there are use cases, in which insights for objects about being clustered or being an outlier is important? In summary, we seek to better understand whether and how visual analytics can benefit from visual approaches combining clustering and outlier analysis to gain a better understanding of the characteristics of a given data set.

This approach is related to other visual analytics approaches that require different classes of algorithmic models. Accordingly, one challenge is the integration of clustering and outlier analysis algorithms in a visual-interactive analysis tool. This challenge comes with the problem of weighting and conflating results of two classes of algorithms in order to better support downstream analysis goals. In addition, both classes of algorithms provide various algorithms with individual parameters. Challenges exist in the visualization of multiple algorithm results as a basis for parameter improvement and the selection of meaningful algorithm results.

Our main contribution is a visual analytics approach that combines the analysis of clusters and outliers, offering four novel ideas. First, with the approach, users have access to clustering as well as outlier analysis techniques and can now simultaneously analyze patterns of clusters and outliers in a visual way. Second, our approach allows to systematically leverage multiple different clustering and outlier algorithms, including the definition and refinement of algorithm parameters. We use dimension reduction as a means to represent high-dimensional data in 2D for any analysis result. Small multiples are used to show different algorithm results side-by-side, fostering visual comparison. Users can now benefit from the characteristics of different implementations, as well as from mitigating the overestimation of single results. Third, we provide a consolidation view, in which a user can interactively integrate and combine the information provided by multiple clustering and outlier analysis algorithms. Towards better supporting users in making informed decisions, visual encodings are provided that resemble the overall algorithm results from different user-selectable weighting strategies. Fourth, in an experiment with handwritten digits data, we demonstrate that the results of cluster and outlier analysis ap-

DOI: 10.2312/eurova.20171114

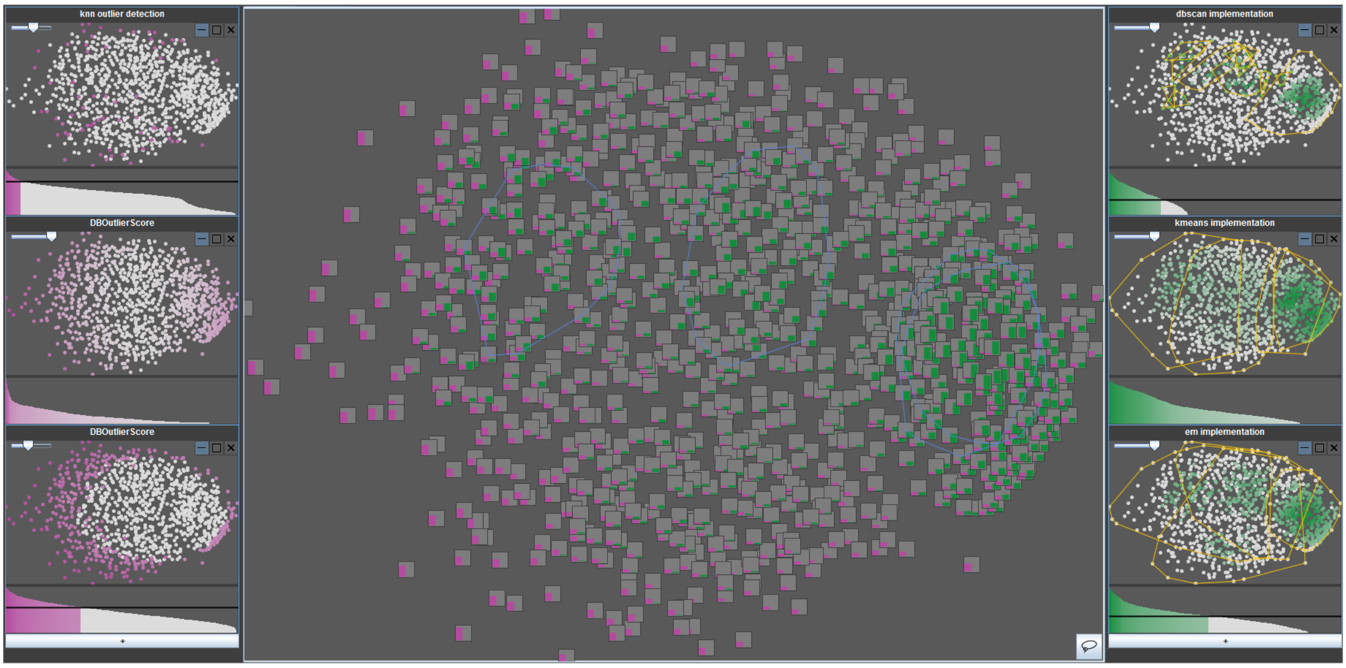

Figure 1: *A visual-interactive system for the combined analysis of clusters and outliers. Three outlier detection algorithms are added to the system (left), as well as three clustering algorithms (right). Visual comparison indicates that algorithm results within both classes can differ considerably. The Consolidation View at the center contains the condensed information of the lateral interfaces, steerable with filtering and weighting interaction. We make two observations. First, cluster and outlier analysis results are not entirely disjoint. Second, three regions are well clustered – and thus used for an interactive labeling task.*

proaches do share common areas of the data set. In a usage scenario using the example of labeling large unknown data collections, we demonstrate that combining both analysis tasks can facilitate the analysis process.

2. Related Work

Related work spans from algorithms for clustering and outlier analysis to visual analytics systems facilitating the interactive clustering and outlier analysis.

2.1. Algorithms for Cluster and Outlier Analysis

A variety of clustering algorithms exist, just as algorithms for outlier analysis. The complexity of the 'design space‘ increases with the parameters for any given algorithm. In this respect, virtually any algorithm has specific characteristics and advantages for specific applications, respectively. In turn, using Anil K. Jain's quote: 'there is no single clustering algorithm that has been shown to dominate other algorithms across all application domains‘ [Jai10]. It can be assumed that the situation is similar with outlier analysis.

Specific classes of clustering algorithms are partitioning-based (e.g., k-means [Mac67]), density-based (e.g., DBScan [EKSX96]), hierarchical (e.g., applied by Seo and Shneiderman [SS02]), or neural network-based (e.g., applied by Schreck et al. [SBVLK09]). We refer to Jain's survey for a recent overview [Jai10]. At a glance, outlier detection algorithms can be differentiated into five different classes, i.e., classification-based, clustering-based, nearest neighbor or density-based [KN98, KShZ08, ZHJ09], statistical [DJC98, Esk00], and spectral (e.g., achieved with dimension reduction [SZS*16]) algorithms. This taxonomy is in line with the survey of Chandola et al. [CBK09].

2.2. Visual-Interactive Cluster and Outlier Analysis

Various visual analytics approaches support the visualization of cluster structures, outliers, or both. In addition, a variety of approaches support defining and steering clustering algorithms while at least a small (disjoint) number of approaches allow steering outlier detection algorithms. For the sake of brevity, we require approaches to allow result visualization, as well as algorithm steering and optimization. Basically, we identified either clustering or outlier approaches. Hardly any approach exists allowing the visual-interactive definition and manipulation of clustering *as well as* outlier algorithms.

A series of approaches supports the visual-interactive definition, manipulation, and analysis of clustering algorithms. Special characteristics are the initialization and visual observation of clustering algorithms [SBVLK09], or the visual comparison of multiple clustering results [LSPS10, BvLBS11, LKS*15] at different granularities. In addition to cluster analysis capability, some approaches also provide visualizations that emphasize anomalies, e.g., with calendar-based views [SBM*14], projection-based visualizations with outlier highlighting [WVZ*15], or with the parallel coordinates technique [LSPS10, LKS*15]. However, approaches for steering clustering algorithms hardly provide algorithms for outlier detection.

Visual analytics approaches involving steering support for algorithms for anomaly detection are scarce. Liao et al. use a conditional random field model to detect anomalies in taxi GPS data

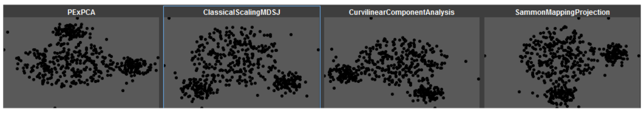

Figure 2: *Choice of a dimension reduction based on visual comparison, using the mouse data set as example. Color coding can be added to assess the quality of representations.*

[LYC10]. In an approach for the detection of spatio-temporal anomalies in large-scale networks, different predefined anomaly detection techniques can be used [LSW13]. Ko et al. support anomaly detection in high-dimensional multivariate network links [KAW*14]. Users can steer the anomaly detection process by selecting (querying) so-called conditional attributes. Finally, the definition and steering of anomaly detection and removal algorithms has been used in a visual-analytics system allowing cleansing time series data [BRG*12]. In summary, these works prefer steering and manipulation of outlier analysis algorithms over cluster analysis tasks.

3. Approach

We present a visual-interactive tool that allows the combined analysis of clustering and outline algorithms. A unified visual-interactive interface supports the selection, parameterization, and analysis of individual algorithms (Section 3.1). To exploit the benefits of different implementations, users can select, compare, and refine multiple clustering and outlier algorithms (Section 3.2). Finally, the Consolidation View condenses relevant information from multiple algorithm results into a single view (Section 3.3).

3.1. Integration of Clustering and Outlier Algorithms

A unified workflow enables the selection, execution, and analysis of clustering and outlier algorithms. A graphical user interface (GUI) allows the parameterization of algorithms from different classes. The dynamically created GUI adapts to parameter sets of individual algorithms, characterizations of parameter types and value ranges enhance the process of parameter tuning. In combination with the algorithm results, the approach uses quality measures to assess cluster and outlier affinity for data objects. Centroid distances for partitioning-based, neural network-based, and hierarchical clusterings, as well as nearest neighbor distances for density-based clusterings are used to assess cluster quality. The output of outlier algorithms can directly be used to assess outlier affinity. Normalizations of individual quality measures produces relative quality scores and facilitates comparability.

Visualization of High-Dimensional Data To achieve a common ground, we use the same visualization technique for any algorithm result. The line of approach reflects the structure of high-dimensional data sets at an elementary level of granularity using dimension reduction [SZS*16] in combination with scatterplots. The linear nature of PCA [Jol02] tends to be sensitive to outliers, which is beneficial for the outlier analysis task. In turn, many non-linear dimension reduction algorithms have advantages in local manifold optimizations, which appeared reduce overplotting challenges. Overall, users can interactively switch between five dimension reduction techniques (PCA [Jol02], MDS [Kru64], Sammons Mapping [Sam69], CCA [DH97], and t-SNE [vdMH08]). Measures like Venna and Kaski's Trustworthiness Measure [VK01] can be used to assess the projection quality. Figure 2 shows a small cutout.

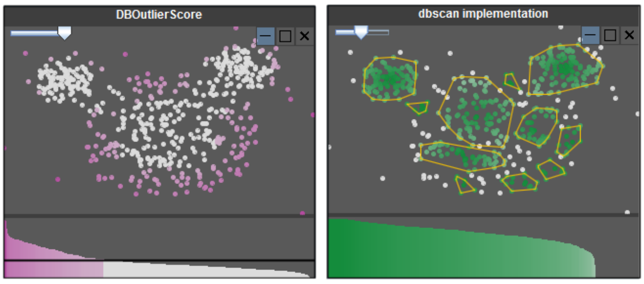

Figure 3: *Visual-interactive result views. Algorithm-dependent quality measures are used to assess the 'interestingness' of every data object. A black horizontal slider at the bottom can be used as a filter control, e.g., to clearly separate outlier objects. At the upper left, users can adjust the weight of the result for the global interestingness scores.*

Visual-Interactive Result Visualization The result of every algorithm is depicted in the visual-interactive Result View showing three core characteristics (Figure 3 depicts two examples). First, the dimensionality-reduced representation of the data is included. Second, color-coding is used for every data object to depict outlier and cluster (quality) scores. A purple-to-gray colormap represents the result of outlier detection algorithms, cluster results are depicted with a green-to-gray colormap. In addition, convex hulls [SP07] encode spatial cluster structures (Figure 3, right). Third, a control at the bottom allows filtering data objects with low scores. The data objects are ranked and aligned with respect to their quality scores from left to right, the encoding of bars shows the distribution of scores in the notion of a pan flute. A horizontal bar serves as the filter control (Figure 3, left).

3.2. Small Multiples of Algorithm Implementations

Users can add multiple clustering and outlier algorithms to exploit the benefit of different algorithm characteristics. We take three requirements into account. First, results of the two algorithm classes are allocated separately, i.e., in two lateral columns for outliers (left) and clusters (right) (cf. Figure 1). Second, different results of the same class are visually comparable in Result Views one below the other. Third, users can refine parameter values to improve algorithms as a result of the visual comparison. To ease the refinement process, the GUI for the selection of algorithms is reused.

3.3. Consolidation of Multiple Algorithm Results

To support informed decision making, users can consolidate the normalized algorithm scores in two ways. First, algorithms results of the same class can be condensed to a single 'interestingness' score. For this purpose, weighting sliders in every result view allow to conflate the strengths of alternative algorithm characteristics (cf. Figure 1). As an alternative, users can switch between objects' minimum, maximum, median, or mean scores achieved with individual algorithms. Second, the two interestingness scores can further be combined, leading to a single variable representing the information of the entire result collection for every data object. In the usage scenario an orange-to-gray colormap is used to depict this overall

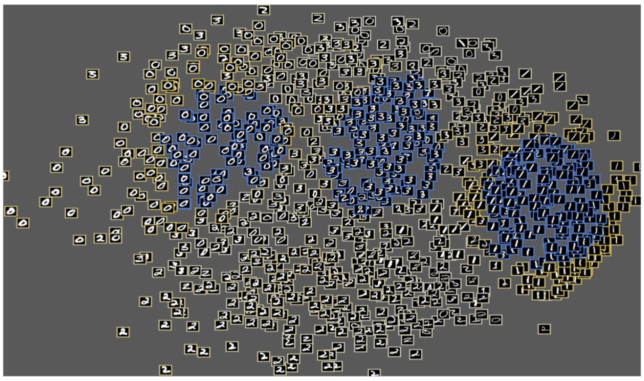

Figure 4: *Use case-specific visualization of handwritten digits. Orange-colored outlines represent the overall score from all cluster and outlier analysis algorithms. Three regions with high orange scores have been used for interactive labeling. The visualization reveals that the selected regions refer to the clearly separated object classes 0, 1, and 3.*

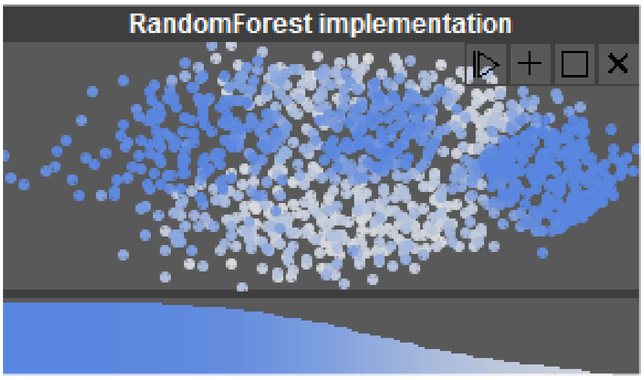

Figure 5: *We use the algorithm selection interface to integrate a classifier into the tool. The least significant confidence measure is used to assess the classification accuracy.*

interestingness score. Figure 4 demonstrates how this information is used to support a class-labeling task.

Consolidation View The Consolidation View (cf. Figure 1, center) shows the results of the two conflation strategies. Bivariate interestingness scores are represented with a small barchart for every data object. In the usage scenario, the global interestingness score is mapped to the outline of every data object using an orange-to-gray colormap. Figure 4 demonstrates the overall scores based on Figure 1.

4. Usage Scenario

We demonstrate the usefulness of the approach using the example of a labeling task. To support labeling, the goal is to identify interesting subsets in the data. A data set of handwritten digits is used, the selected subset is considered unknown at start. We add three outlier analysis and three clustering algorithms. Parameter tuning, filtering, and weighting of algorithms leads to the state presented in Figure 1. We make three observations. First, the approach clarifies that results of the same algorithm class reveal considerable differences, for both outlier analysis and clustering. Second, the small bar charts at the center show that some data objects have an affinity to both being an outlier and a cluster member. This brings interesting insight regarding the question whether outlier and cluster areas are necessarily disjoint. Finally, we identify three cluster regions. For the labeling, we group each cluster region with a lasso tool (thin blue lines).

Figure 4 shows the semantical perspective of the data set, consisting of the handwritten digits 0, 1, 2, and 3. It can be seen that the distribution of handwritten digits objects on the manifold matches the three previously labeled regions colored blue. Orange color depicts the overall interestingness scores for unlabeled data (cf. Section 3.3). With an increasing number of labels, outliers become more interesting for the labeling task (see, e.g., the orange outliers on the left), referring to classification problems for critical cases. Thus, we take the labeling task one step further and train classifier. In combination with a quality measure (least significant confidence [Set09]), the labeling task turns into an 'Interactive Learning' [HNH*12] scenario. The classification result is demonstrated in Figure 5. The classifier has high accuracy near the labels as well as on the left-hand side. However, especially between clusters additional labels will be needed to increase the classifier performance. In particular, outlier analysis results can help to label critical cases.

5. Conclusion

We presented a tool that combines cluster and outlier analysis. Users are able to combine multiple algorithms of both classes and thus can condense the information provided by different implementations to make more informed decisions. The results of a usage scenario demonstrates that the approach is useful, e.g., for labeling large unknown data sets. Future work includes experiments supporting other analysis tasks apart from labeling. In this connection, we draw the connection to approaches facilitating relation-seeking tasks, e.g. to discover interesting relations between clusters and metadata [BRS*12, BSW*14]. In addition, parameter space analysis and user guidance can further extend the approach. In general, we see our work in line with the emerging field of Progressive Visual Analytics (PVA). While others have focused on making algorithmic process more transparent [MPG*14] and on quality metrics for steering the visualization process [SASS16], our work focuses on the combination and integration of different analysis tasks. We hope that others will be inspired by our work and follow up with further research on the exiting area of PVA.

6. Acknowledgments

We thank the Machine Learning Group at the University of Waikato for providing the WEKA [SF16], the Database Systems Group at Ludwig Maximilian University of Munich for providing the ELKI [SKE*15], and the Algorithmics Group at the University of Konstanz in Germany for providing the MDSJ library. This work was partly supported by the Vienna Business Agency.

References

[AA06] ANDRIENKO N., ANDRIENKO G.: *Exploratory Analysis of Spatial and Temporal Data: A Systematic Approach*. Springer-Verlag New York, Inc., 2006. doi:10.1007/3-540-31190-4. 1

[BRG*12] BERNARD J., RUPPERT T., GOROLL O., MAY T., KOHLHAMMER J.: Visual-interactive preprocessing of time series data.

In *Proc. of EG SIGRAD* (2012), vol. 81, Linköping University Electronic Press, pp. 39–48. 3

[BRS*12] Bernard J., Ruppert T., Scherer M., Schreck T., Kohlhammer J.: Guided discovery of interesting relationships between time series clusters and metadata properties. In *Knowledge Management and Knowledge Technologies (i-KNOW)* (2012), ACM, pp. 22:1–22:8. doi:10.1145/2362456.2362485. 4

[BSW*14] Bernard J., Steiger M., Widmer S., Lücke-Tieke H., May T., Kohlhammer J.: Visual-interactive Exploration of Interesting Multivariate Relations in Mixed Research Data Sets. *Computer Graphics Forum (CGF) 33*, 3 (2014), 291–300. doi:10.1111/cgf.12385. 4

[BvLBS11] Bernard J., von Landesberger T., Bremm S., Schreck T.: Multiscale visual quality assessment for cluster analysis with Self-Organizing Maps. In *SPIE Visualization and Data Analysis (VDA)* (2011), SPIE Press, pp. 78680N.1 – 78680N.12. doi:10.1117/12.872545. 2

[CBK09] Chandola V., Banerjee A., Kumar V.: Anomaly detection: A survey. *ACM Comput. Surv. 41*, 3 (2009), 15:1–15:58. doi:10.1145/1541880.1541882. 1, 2

[DH97] Demartines P., Herault J.: Curvilinear component analysis: a self-organizing neural network for nonlinear mapping of data sets. *Neural Networks, IEEE Transactions on 8*, 1 (1997), 148–154. doi:10.1109/72.554199. 3

[DJC98] Desforges M. J., Jacob P. J., Cooper J. E.: Applications of probability density estimation to the detection of abnormal conditions in engineering. In *Institute of Mechanical Engineers* (1998), vol. 212, pp. 687–703. 2

[EKSX96] Ester M., Kriegel H.-P., Sander J., Xu X.: A density-based algorithm for discovering clusters in large spatial databases with noise. AAAI Press, pp. 226–231. 2

[Esk00] Eskin E.: Anomaly detection over noisy data using learned probability distributions. In *Machine Learning (ICML)* (2000), Morgan Kaufmann Publishers Inc., pp. 255–262. 2

[HNH*12] Höferlin B., Netzel R., Höferlin M., Weiskopf D., Heidemann G.: Inter-active learning of ad-hoc classifiers for video visual analytics. In *Visual Analytics Science and Technology (VAST)* (2012), IEEE, pp. 23–32. doi:10.1109/VAST.2012.6400492. 4

[Jai10] Jain A. K.: Data clustering: 50 years beyond k-means. *Pattern Recognition Letters 31*, 8 (2010), 651–666. 1, 2

[Jol02] Jolliffe I. T.: *Principal Component Analysis*, 3rd ed. Springer, 2002. 3

[KAW*14] Ko S., Afzal S., Walton S., Yang Y., Chae J., Malik A., Jang Y., Chen M., Ebert D.: Analyzing high-dimensional multivariate network links with integrated anomaly detection, highlighting and exploration. In *IEEE (VAST)* (2014), pp. 83–92. doi:10.1109/VAST.2014.7042484. 3

[KN98] Knorr E. M., Ng R. T.: Algorithms for mining distance-based outliers in large datasets. In *Very Large Data Bases (VLDB)* (San Francisco, CA, USA, 1998), Morgan Kaufmann Publishers Inc., pp. 392–403. 2

[Kru64] Kruskal J. B.: Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika 29*, 1 (1964), 1–27. 3

[KShZ08] Kriegel H.-P., S hubert M., Zimek A.: Angle-based outlier detection in high-dimensional data. In *ACM SIGKDD Knowledge Discovery and Data Mining (KDD)* (2008), ACM, pp. 444–452. doi:10.1145/1401890.1401946. 2

[LKS*15] L'Yi S., Ko B., Shin D., Cho Y.-J. J., Lee J., Kim B., Seo J.: XCluSim: a visual analytics tool for interactively comparing multiple clustering results of bioinformatics data. *BMC bioinformatics 16 Suppl 11* (2015). 2

[LSPS10] Lex A., Streit M., Partl C., Schmalstieg D.: Comparative analysis of multidimensional, quantitative data. *IEEE Transactions on Visualization and Comp. Graphics 16*, 6 (2010), 1027–1035. doi:10.1109/TVCG.2010.138. 2

[LSW13] Liao Q., Shi L., Wang C.: Visual analysis of large-scale network anomalies. *IBM Journal of R&D 57*, 3/4 (2013), 13:1–13:12. doi:10.1147/JRD.2013.2249356. 3

[LYC10] Liao Z., Yu Y., Chen B.: Anomaly detection in GPS data based on visual analytics. In *Visual Analytics Science and Technology (VAST)* (2010), IEEE Computer Society, pp. 51–58. doi:10.1109/VAST.2010.5652467. 3

[Mac67] MacQueen J. B.: Some methods for classification and analysis of multivariate observations. In *Berkeley Symposium on Mathematical Statistics and Probability* (1967), vol. 1, University of California Press, pp. 281–297. 2

[MPG*14] Mühlbacher T., Piringer H., Gratzl S., Sedlmair M., Streit M.: Opening the black box: Strategies for increased user involvement in existing algorithm implementations. *IEEE TVCG 20*, 12 (2014), 1643–1652. doi:10.1109/TVCG.2014.2346578. 4

[Sam69] Sammon J. W.: A nonlinear mapping for data structure analysis. *IEEE Trans. Comput. 18*, 5 (1969), 401–409. doi:10.1109/T-C.1969.222678. 3

[SASS16] Schulz H. J., Angelini M., Santucci G., Schumann H.: An enhanced visualization process model for incremental visualization. *IEEE Transactions on Visualization and Computer Graphics 22*, 7 (2016), 1830–1842. doi:10.1109/TVCG.2015.2462356. 4

[SBM*14] Steiger M., Bernard J., Mittelstädt S., Lücke-Tieke H., Keim D., May T., Kohlhammer J.: Visual analysis of time-series similarities for anomaly detection in sensor networks. *Computer Graphics Forum (CGF) 33*, 3 (2014), 401–410. doi:10.1111/cgf.12396. 2

[SBVLK09] Schreck T., Bernard J., Von Landesberger T., Kohlhammer J.: Visual cluster analysis of trajectory data with interactive kohonen maps. *Information Visualization 8*, 1 (2009), 14–29. doi:10.1057/ivs.2008.29. 2

[Set09] Settles B.: *Active Learning Literature Survey*. Tech. Report 1648, Univ. of Wisconsin–Madison, 2009. 4

[SF16] Smith T. C., Frank E.: *Statistical Genomics: Methods and Protocols*. Springer, New York, NY, 2016, ch. Introducing Machine Learning Concepts with WEKA, pp. 353–378. 4

[SKE*15] Schubert E., Koos A., Emrich T., Züfle A., Schmid K. A., Zimek A.: A framework for clustering uncertain data. *PVLDB 8*, 12 (2015), 1976–1979. 4

[SP07] Schreck T., Panse C.: A new metaphor for projection-based visual analysis and data exploration. In *Visualization and Data Analysis* (2007), no. 6495 in SPIE Proceedings, SPIE, pp. 64950L–64950L–12. doi:10.1117/12.697879. 3

[SS02] Seo J., Shneiderman B.: Interactively exploring hierarchical clustering results. *Computer 35*, 7 (2002), 80–86. doi:10.1109/MC.2002.1016905. 2

[SZS*16] Sacha D., Zhang L., Sedlmair M., Lee J. A., Peltonen J., Weiskopf D., North S. C., Keim D. A.: Visual Interaction with Dimensionality Reduction: A Structured Literature Analysis. *IEEE TVCG 23*, 01 (2016), 241–250. doi:10.1109/TVCG.2016.2598495. 2, 3

[vdMH08] van der Maaten L., Hinton G. E.: Visualizing high-dimensional data using t-sne. *Journal of Machine Learning Research 9* (2008), 2579–2605. 3

[VK01] Venna J., Kaski S.: Neighborhood preservation in nonlinear projection methods: An experimental study. In *Artificial Neural Networks (ICANN)* (2001), Springer-Verlag, pp. 485–491. 3

[WVZ*15] Wilhelm N., Vögele A., Zsoldos R., Licka T., Krüger B., Bernard J.: Furyexplorer: visual-interactive exploration of horse motion capture data. In *SPIE, Visualization and Data Analysis (VDA)* (2015), pp. 93970F–93970F–15. doi:10.1117/12.2080001. 2

[ZHJ09] Zhang K., Hutter M., Jin H.: A new local distance-based outlier detection approach for scattered real-world data. In *Pacific-Asia Conference on KDD (PAKDD)* (2009), pp. 813–822. doi:10.1007/978-3-642-01307-2_84. 2