Eurographics Conference on Visualization (EuroVis), Posters Track (2017)
A. Puig Puig and T. Isenberg (Editors)

Integrating Guided Clustering in Visual Analytics to Support Domain Expert Reasoning Processes

Andreas Mathisen[1], Matthias Nielsen [2] and Kaj Grønbæk[1]

[1] Aarhus University, Denmark, [2] The Alexandra Institute

Abstract

Recent research shows promise in combining Information Visualization (IV) and Machine Learning (ML) to assist data analysis performed by domain experts. However, this approach presents non-trivial challenges, in particular when the goal is to incorporate knowledge provided by the domain expert in underlying ML algorithms. To address these challenges, we present an analytical process and a visual analytics tool that uses visual queries to capture examples from the domain experts' existing reasoning process which will guide the subsequent clustering. Our work is motivated by a collaboration with personnel at the Danish Business Authority, who are interested in two types of insights: (1) On which data dimensions is a selected subset of companies different from the remaining companies? (2) Which other companies lie within the same multi-dimensional subspace? The poster will illustrate a real analysis scenario, where the presented analytic process allows auditors to use their knowledge of identified "suspicious" companies to kick-start the analysis for others.

1. Introduction

Combining IV and ML was recently suggested as being a core research objective at a Dagstuhl Seminar [KMRV15], to extend the existing work on using ML methods within visual environments. Numerous approaches have been introduced to visually convey high-dimensional data, for instance using lower dimensional projections or clustering algorithms. However, applying ML algorithms in practice is usually an iterative process, where the designer extracts new features and validates intermediate results. Since this process can be challenging, it typically requires domain expert knowledge. We present an analytical approach that exploits the scenario in which domain experts can provide a partial labeling, i.e. instances of interest to their analysis. The core idea is to find relevant clustering results using a two-round clustering approach guided by examples which domain experts can provide via visual queries.

Our work is based on a collaboration with personnel at the Danish Business Authority, who lack automated tools to systematically exploit their data to, e.g., uncover fraudulent behavior. We found that their analytical reasoning processes are often started from examples or risk factors derived from previous cases (e.g. bankrupt companies). Given the nature of available examples the resulting labeling of the companies is only partial which can be challenging to cope with in ML. Concretely, we found that the knowledge provided by the auditors suffers from two distinct characteristics, which we denote *abstract* and *incomplete*. A labeling is *abstract* w.r.t. label A if the items labeled as A are not similar in the feature space and therefore should have sub-labels, as illustrated with different decision boundaries in fig. 1a. A labeling is *incomplete* w.r.t. label A if further instances should have label A additional to those currently labeled as A, as illustrated in fig. 1b. Intuitively, these additional instances are of utmost interest, since they are similar to the provided examples in the feature space. Note, that a labeling can both be abstract and incomplete w.r.t. a label A, and if this is the case it can be difficult to find satisfactory results with conventional supervised or semi-supervised learning methods.

(a) *Abstract labeling.* **(b)** *Incomplete labeling.*

Figure 1: *Properties of partial labeling.*

2. Related Work

Analysing high-dimensional data is an active field of research within both the IV community and the ML community. Liu et al. [LMW*16] recently provided a thorough review of the recent advances in high-dimensional data visualization. Several techniques exist for visual mapping of multiple dimensions [EDF08, ID90, NG15, Kan00, FCI05] as well as visualizing uncertainty [CCM13]. However, visualizing all dimensions severely limits our ability to spot meaningful patterns. A common approach is therefore to project high-dimensional data to lower dimensional spaces to enable simpler visual mappings [JZF*09, LWBP14, CLKP10]. Vi-

DOI: 10.2312/eurp.20171164

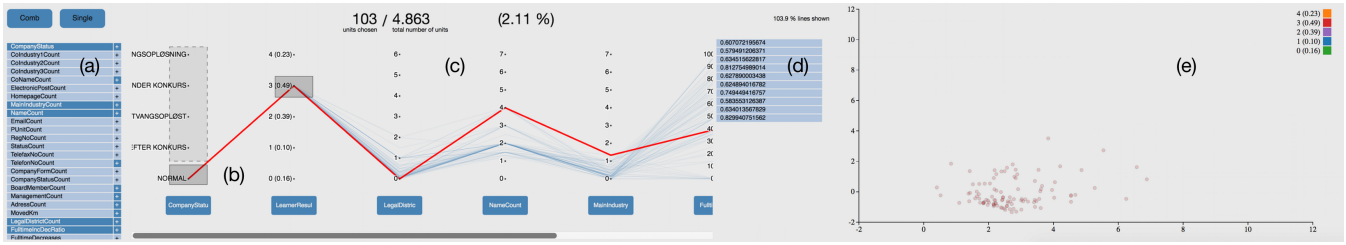

Figure 2: *A visual analytics tool where a parallel coordinates visualization (c) is enhanced with clustering functionality. Users can (a) select features of interest and (b) provide visual queries using brushes to the clustering process. Afterwards users can inspect the best results shown with the V-measures (d) using two coordinated views (c) and (e).*

sual tools have also been used to inspect ML results [FWR99] in order to understand the output or to manipulate the model [GXWY10, BLBC12]. The visual analytics concept is excellent to support exploratory analysis that incorporates domain knowledge [SVW*10] and various approaches have been proposed to achieve this goal [HDK*07, HBM*13, Gle13].

3. Exploiting Domain Knowledge

To exploit domain knowledge that is *abstract* and *incomplete* we propose an analytical process consisting of three steps: (1) *define examples*, (2) *generate clusters* and (3) *inspect results*. In our prototype, we use conventional methods to visualize high-dimensional data; parallel coordinates [ID90, NG15, FWR99] for the multidimensional features space and scattersplots for the reduced feature space. Figure 2 depicts the web-based prototype with two coordinated views that displays one of the potential clustering results.

(1) Define examples: The user can provide examples using visual queries (brushing in our case [BC87, HS04]) in the parallel coordinates visualization, which then generates a binary distinction. The instances satisfying the current selections are one group and the remaining instances constitute the other group. This allows to effectively compare the selected examples with the rest. The user can furthermore choose to limit the feature space by selecting only those features of interest to the current analysis.

(2) Generate clusters: A two-round clustering is utilized based on the visual query of a user. In the first round, clustering is performed on each initial group of instances defined by the user's query. In this round we use the silhouette coefficient [Rou87] to reason about the structural properties of the clusters to find the optimal number. The result of the first round is a sub-labeling of the examples, i.e. it is a way to deal with an *abstract* labeling. In the second round, clustering is performed on the entire data set to deal with an *incomplete* labeling. In this round we use combinations of the sub-labels found in the first round together with the V-measure [RH07] to find the optimal parameters. While our method is not specific to a single clustering algorithm, we use the K-means clustering algorithm [AV07] due to its speed. We search for results both in the number of clusters and in the feature space, and continuously report the best results found so far. To verify the usefulness of our process, we applied it also to the popular Iris data set [Lic13]. The Iris data set contains 3 classes, but using clustering on this data set will traditionally yield only 2 clusters. However, if an expert can provide a partial labeling which separates the majority of the two similar classes, our approach will suggest 3 clusters.

(3) Inspect results: The clustering results will be presented as a new axis in the parallel coordinates visualization and color-coded in the scatterplot, where the PCA algorithm [TB99] is used to reduce the feature space. The views are coordinated, so users can update both views by either hovering the scatterplot or by creating filters in the parallel coordinates visualization.

4. Applied to the Business Auditing Case

The motivating use case for this analytical approach is to support business audit personnel in identifying fraudulent or otherwise troublesome companies. Currently, the selection of which companies to investigate is based on whether individual companies satisfy some of the known risk factors, using either historical registration data (e.g. board members), employment data or financial data. As an example, we converted the registration data to features by counting the number of occurrences for each type of registration. We then normalized the resulting data with the time span between the first and last occurrence. The data presented in Figure 2 shows the companies in Denmark with the most registration updates. In the example in Figure 2 all companies with a status different from normal are queried as one class. The resulting labeling is *abstract*, since the status does not describe why a company has gone bankrupt or been forced to dissolute. From this example we for instance learned that if a company changes name more frequently than business type and legal district, they are within a cluster where 100/202 of the companies have stopped. Since we believe the labeling to also be *incomplete*, we interpret the 102 remaining companies to be more suspicious than a random one out of all the 3836 normal companies.

5. Conclusion

In real world scenarios it is infeasible to expect perfect domain information, hence we have presented an approach that can still utilize partial information in the underlying clustering process. We present a prototype tool that incorporates our analytical approach and we provide a proof of concept of our approach in a relevant use case. Immediate future work include enhancing the usability of our prototype by doing additional user studies with the Business Audit personnel. We will also investigate how to mitigate potential expectation or confirmation biases, which can be prominent when inexperienced users are evaluating ML results.

Acknowledgements

This work was conducted in the DABAI project (IFD-5153-00004B) supported by the Innovation Fund Denmark.

References

[AV07] ARTHUR D., VASSILVITSKII S.: k-means++: The advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms* (2007), Society for Industrial and Applied Mathematics, pp. 1027–1035. 2

[BC87] BECKER R. A., CLEVELAND W. S.: Brushing scatterplots. *Technometrics 29*, 2 (1987), 127–142. 2

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

[CCM13] CHAN Y.-H., CORREA C. D., MA K.-L.: The generalized sensitivity scatterplot. *IEEE transactions on visualization and computer graphics 19*, 10 (2013), 1768–1781. 1

[CLKP10] CHOO J., LEE H., KIHM J., PARK H.: ivisclassifier: An interactive visual analytics system for classification based on supervised dimension reduction. In *Visual Analytics Science and Technology (VAST), 2010 IEEE Symposium on* (2010), IEEE, pp. 27–34. 1

[EDF08] ELMQVIST N., DRAGICEVIC P., FEKETE J.-D.: Rolling the dice: Multidimensional visual exploration using scatterplot matrix navigation. *IEEE transactions on Visualization and Computer Graphics 14*, 6 (2008), 1539–1148. 1

[FCI05] FANEA E., CARPENDALE S., ISENBERG T.: An interactive 3d integration of parallel coordinates and star glyphs. In *Information Visualization, 2005. INFOVIS 2005. IEEE Symposium on* (2005), IEEE, pp. 149–156. 1

[FWR99] FUA Y.-H., WARD M. O., RUNDENSTEINER E. A.: Hierarchical parallel coordinates for exploration of large datasets. In *proc. of the conference on Visualization'99: celebrating ten years* (1999), IEEE Computer Society Press, pp. 43–50. 2

[Gle13] GLEICHER M.: Explainers: Expert explorations with crafted projections. *IEEE transactions on visualization and computer graphics 19*, 12 (2013), 2042–2051. 2

[GXWY10] GUO P., XIAO H., WANG Z., YUAN X.: Interactive local clustering operations for high dimensional data in parallel coordinates. In *Visualization Symposium (PacificVis), 2010 IEEE Pacific* (2010), IEEE, pp. 97–104. 2

[HBM*13] HU X., BRADEL L., MAITI D., HOUSE L., NORTH C.: Semantics of directly manipulating spatializations. *IEEE Transactions on Visualization and Computer Graphics 19*, 12 (2013), 2052–2059. 2

[HDK*07] HAO M. C., DAYAL U., KEIM D. A., MORENT D., SCHNEIDEWIND J.: Intelligent visual analytics queries. In *Visual Analytics Science and Technology, 2007. VAST 2007. IEEE Symposium on* (2007), IEEE, pp. 91–98. 2

[HS04] HOCHHEISER H., SHNEIDERMAN B.: Dynamic query tools for time series data sets: timebox widgets for interactive exploration. *Information Visualization 3*, 1 (2004), 1–18. 2

[ID90] INSELBERG A., DIMSDALE B.: Parallel coordinates: a tool for visualizing multi-dimensional geometry. In *Proceedings of the 1st conference on Visualization'90* (1990), IEEE Computer Society Press, pp. 361–378. 1, 2

[JZF*09] JEONG D. H., ZIEMKIEWICZ C., FISHER B., RIBARSKY W., CHANG R.: ipca: An interactive system for pca-based visual analytics. In *Computer Graphics Forum* (2009), vol. 28, Wiley Online Library, pp. 767–774. 1

[Kan00] KANDOGAN E.: Star coordinates: A multi-dimensional visualization technique with uniform treatment of dimensions. In *Proceedings of the IEEE Information Visualization Symposium* (2000), vol. 650, Citeseer, p. 22. 1

[KMRV15] KEIM D. A., MUNZNER T., ROSSI F., VERLEYSEN M.: Bridging information visualization with machine learning (dagstuhl seminar 15101). *Dagstuhl Reports 5*, 3 (2015). 1

[Lic13] LICHMAN M.: UCI machine learning repository, 2013. URL: http://archive.ics.uci.edu/ml. 2

[LMW*16] LIU S., MALJOVEC D., WANG B., BREMER P.-T., PASCUCCI V.: Visualizing high-dimensional data: Advances in the past decade. *IEEE Transactions on Visualization and Computer Graphics* (2016). 1

[LWBP14] LIU S., WANG B., BREMER P.-T., PASCUCCI V.: Distortion-guided structure-driven interactive exploration of high-dimensional data. In *Computer Graphics Forum* (2014), vol. 33, Wiley Online Library, pp. 101–110. 1

[NG15] NIELSEN M., GRØNBÆK K.: Pivotviz: Interactive visual analysis of multidimensional library transaction data. In *proc. of the 15th ACM/IEEE-CS Joint Conference on Digital Libraries* (2015), ACM, pp. 139–142. 1, 2

[RH07] ROSENBERG A., HIRSCHBERG J.: V-measure: A conditional entropy-based external cluster evaluation measure. In *EMNLP-CoNLL* (2007), vol. 7, pp. 410–420. 2

[Rou87] ROUSSEEUW P. J.: Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics 20* (1987), 53–65. 2

[SVW*10] SHRINIVASAN Y. B., VAN WIJK J., ET AL.: Supporting exploratory analysis with the select & slice table. In *Computer Graphics Forum* (2010), vol. 29, Wiley Online Library, pp. 803–812. 2

[TB99] TIPPING M. E., BISHOP C. M.: Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 61*, 3 (1999), 611–622. 2