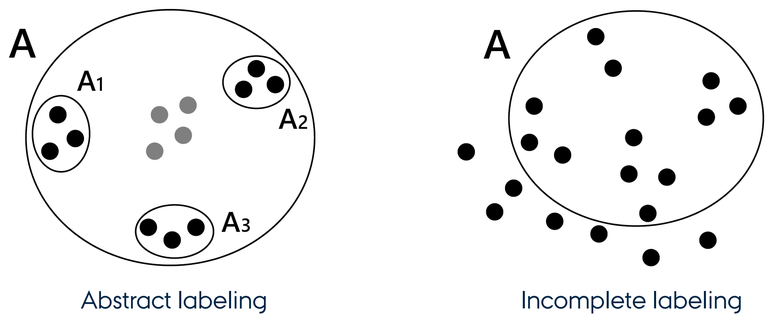
Integrating Guided Clustering in Visual Analytics to Support Domain Expert Reasoning Processes

Andreas Mathisen[1], Matthias Nielsen[2] and Kaj Grønbæk[1]
[1]Aarhus University, Denmark, [2]The Alexandra Institute

Supporting Domain Expert Reasoning

When combining Information Visualization (IV) and Machine Learning (ML) to assist data analysis conducted by domain experts the goal is often to leverage the domain knowledge in the underlying ML algorithms. We present an analytical process and a visual analytics tool that uses visual queries to capture examples from the domain experts' existing reasoning process to guide the subsequent clustering. In collaboration with personnel at the Danish Business Authority, we found that their analytical reasoning processes often start with examples or risk factors derived from previous cases. Given the nature of the 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:

- 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.
- A labeling is *incomplete* w.r.t. label A if further instances should have label A additional to those currently labeled as A.



Abstract labeling | Incomplete labeling

Guided Clustering using Domain Knowledge

We have developed a prototype tool that incorporates the following process, where *abstract* and *incomplete* domain knowledge is used in the data clustering:

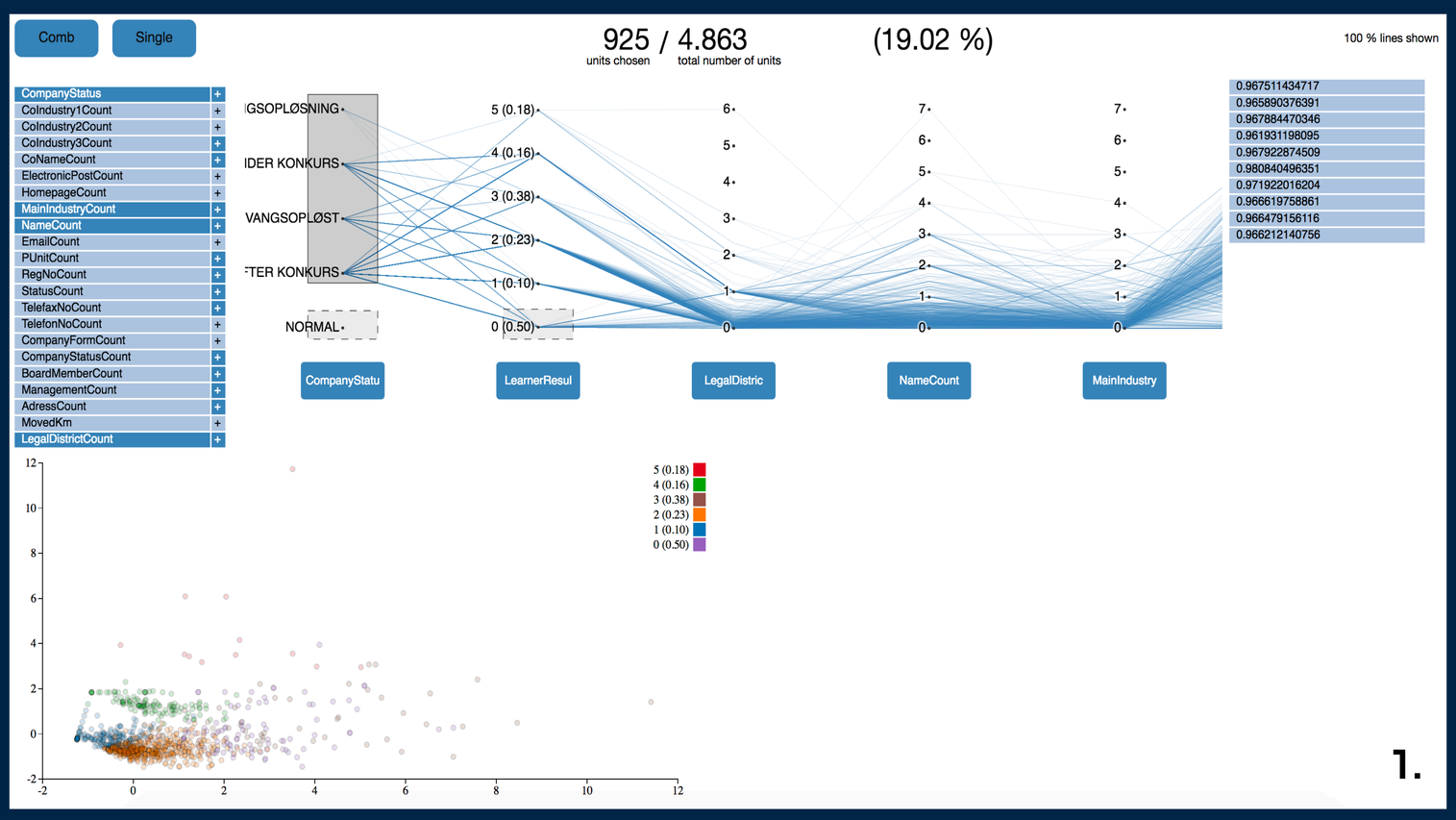
I. **Define examples:** Visual queries (i.e. brushes) in the parallel coordinates visualization can be used to generate a binary labeling.

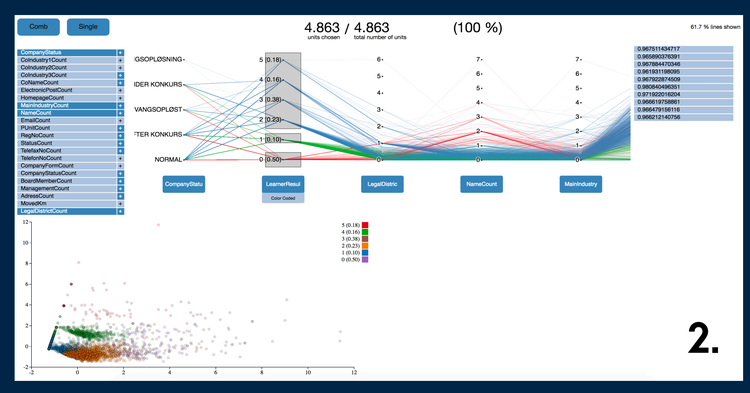
II. **Generate clusters:** We use the K-means algorithm [1] due to its speed, however, this approach is not limited to a single clustering algorithm.
   I. Clustering is performed on each initial group of instances defined by the user' query. In this round we use the silhouette coefficient [2] to reason about the structural properties of the clusters to find the optimal number, to deal with an *abstract* labeling.
   II. 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 [3] to find the optimal parameters.

III. **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 is used to reduce the feature space.

Verification with the Iris Dataset

To verify the usefulness of our process, we applied it also to the popular Iris data set. 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.





Use Case: Business Auditing

Current investigations are based on whether individual companies satisfy some of the known risk factors based on e.g. registration and employment data. In our study we generated additional features by e.g.:

- Counting occurrences for each type of registration.
- Normalizing the counts with the time span between the first and last occurrence.

Analysis example:

- Figure 1, 2 and 3 shows the companies in Denmark with the most registration updates.
- Companies with a status different from normal are queried as one *abstract* class.
- Outcome: 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 the labeling is *incomplete*, we interpret the 102 remaining companies to be more suspicious than a random one out of all the 3836 normal companies.

1. The active brush indicates the current binary labeling. The resulting clusters are augmented with the percentage of the interesting instances situated in a particular cluster. To the right are the potential results displayed together with the V-measure.

2. Clusters can be compared with color-coding in the parallel coordinates visualization.

3. Individual instances can be identified across the views and selected for further investigation.

[1] ARTHUR D., V ASSILVITSKII 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] 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.
[3] ROSENBERG A., H IRSCHBERG J.: V-measure: A conditional entropy-based external cluster evaluation measure. In EMNLP-CoNLL (2007), vol. 7, pp. 410–420.

AARHUS UNIVERSITY
DEPARTMENT OF COMPUTER SCIENCE

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

SOLIDUM PETIT IN PROFUNDIS · UNIVERSITAS ARHUSIENSIS