PR1/2 Demo

Christian Permann

Faculty of Computer Science, University of Vienna,
Währinger Straße 29, 1090 Vienna

13.11.2018

Milestones P1 updated

- Create a basic visualization tool. (done)
- Define interface for data manipulation. (done)
- (Be compatible with ELKI.)
- Allow data import. (CSV/Arff available, done)
- Create data generation logic. (ELKI generator available, possibly add simple GUI with reduced functionality)
- Create dim-reduction for visualization and implement it. (PCA and T-SNE, done)
- Allow to generate a scatter plot matrix. (done)

The Tool

Demo

The Tool - Dimensionality Reduction Interface

```java
public interface IDimensionalityReduction {
    JPanel getOptionsPanel();

    String getName();

    boolean reduce(PointContainer container);
}
```

Figure 1: The interface for dimensionality reduction

Papers for PR2

- LineUp: Visual Analysis of Multi-Attribute Rankings [1]
- WeightLifter: Visual Weight Space Exploration for Multi-Criteria Decision Making [2]
- Metric Factorization for Exploratory Analysis of Complex Data [3]
- DimStiller: Workflows for dimensional analysis and reduction [4]
- Comparing clusterings: an axiomatic view [5]
- Comparing subspace clusterings [6]
- External evaluation measures for subspace clustering [7]

Ideas for evaluating Clusterings

- different quality measures as described in [7]
- a weighted average of measures like in ClusterVision[8]
- using [1, 2, 3] as basis for analyzing the quality of measures and deciding on different weights (kind of a better reasoning for the choice compared to clustervision)
- **clustering (OPTICS?) clusterings and visualizing groups of results across multiple algorithms and settings (new idea?) using [5, 6]** in regards to the distance measure
- possibly training a weighted average of measures via supervision with a Neural Network (needs lots of data and known optimal clusterings; generalizeable result?)

My Idea

Using OPTICS to cluster clusterings could result in a reachability plot that hierarchically shows groups of clustering results that agree on the result. The output of the algorithm can also be visualized in a symmetric heat-map as shown in our OPTICSVis Project. Here it may be possible to see which clusterings overlap with others and which are vastly different, in a hierarchical view. One problem here may be though, that the distance measure should satisfy the triangle inequality for useful measures(, I think). (see Figure 2: Clustering Error Metric)

Distance Measure

3.3.1 Clustering Error

Consider subspace clusterings $\mathcal{S} = \{S_1, S_2, \ldots, S_K\}$ and $\mathcal{S}' = \{S'_1, S'_2, \ldots, S'_{K'}\}$ of K and K' clusters, respectively. Recall from Section 2 that a confusion matrix $M = (M_{ij})$ is a $K \times K'$ matrix in which m_{ij} is the number of data matrix elements shared by the clusters S_i and S'_j. More formally, $m_{ij} = |\mathrm{supp}\,(S_i) \cap \mathrm{supp}\,(S'_j)|$. Note, however, that in the case of subspace clusters, the rows and the columns of M do not necessarily sum up to the cluster sizes. That is, $\sum_i m_{ij} \leq |S'_j|$ and $\sum_j m_{ij} \leq |S_i|$.

Let us transform M into a square matrix by adding rows or columns of zeroes if necessary and use the Hungarian method [43] to find a permutation of the cluster labels such that the sum of the diagonal elements of M is maximized. Denote this maximized sum by D_{max}. Now, we define the clustering error (CE) for subspace clusterings as

$$\mathrm{CE}(\mathcal{S}, \mathcal{S}') = \frac{|U| - D_{max}}{|U|}. \tag{3}$$

In the case of ordinary clusterings (partitions of the rows of the data matrix), the clustering error defined here is the clustering error of Section 2.

Figure 2: Distance Measure from [6]

S. Gratzl, A. Lex, N. Gehlenborg, H. Pfister, and M. Streit, "Lineup: Visual analysis of multi-attribute rankings," *IEEE Transactions on Visualization and Computer Graphics (InfoVis '13)*, vol. 19, no. 12, pp. 2277–2286, 2013.

S. Pajer, M. Streit, T. Torsney-Weir, F. Spechtenhauser, T. Möller, and H. Piringer, "Weightlifter: Visual weight space exploration for multi-criteria decision making," *IEEE Transactions on Visualization and Computer Graphics*, vol. 23, pp. 611–620, January 2017.

C. Plant, "Metric factorization for exploratory analysis of complex data," in *2014 IEEE International Conference on Data Mining*, pp. 510–519, Dec 2014.

S. Ingram, T. Munzner, V. Irvine, M. Tory, S. Bergner, and T. Möller, "Dimstiller: Workflows for dimensional analysis and reduction," in *IEEE Conference on Visual Analytics in Science and Technology (VAST)*, pp. 3–10, October 2010.
(26 out of 94 accepted).

M. Meilă, "Comparing clusterings: an axiomatic view," in *In ICML '05: Proceedings of the 22nd international conference on Machine learning*, pp. 577–584, ACM Press, 2005.

A. Patrikainen and M. Meila, "Comparing subspace clusterings," *IEEE Transactions on Knowledge and Data Engineering*, vol. 18, pp. 902–916, July 2006.

S. Günnemann, I. Färber, E. Müller, I. Assent, and T. Seidl, "External evaluation measures for subspace clustering," in *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, CIKM '11, (New York, NY, USA), pp. 1363–1372, ACM, 2011.

B. C. Kwon, B. Eysenbach, J. Verma, K. Ng, C. D. Filippi, W. F. Stewart, and A. Perer, "Clustervision: Visual supervision of unsupervised clustering," *IEEE Transactions on Visualization and Computer Graphics*, vol. 24, pp. 142–151, Jan 2018.