Explorative Analysis and Visualization of Large Information Spaces

Learning in Parallel Universes

Bernd Wiswedel

Universität Konstanz

1 Research Topic

Learning in Parallel Universes is a relatively new data mining research field; it refers to the setup of having different descriptor spaces in parallel, whereby each universe typically reflects different properties of the underlying objects. The learning task itself is two-fold: to exploit these multiple object representations and link structures that are shared among multiple – though not necessarily all – universes and, secondly, to also identify structures that are specific to individual universes (due to the inherent nature of the respective universes). A related learning task is the so-called Multi-View-Learning, however it assumes all universes (i. e. views) to contain the same structures, whereby the focus is on maximizing consensus between the universes.
The thesis deals with the formalization of learning in Parallel Universes and the development of methods for both supervised and unsupervised problems.

2 Applications

Parallel Universes are essentially encountered in all domains which relate to the computer driven analysis of complex objects, e. g.

Molecular data Molecules are described based on, among others, (i) 3D layout, (ii) charge distribution, (iii) connectivity information

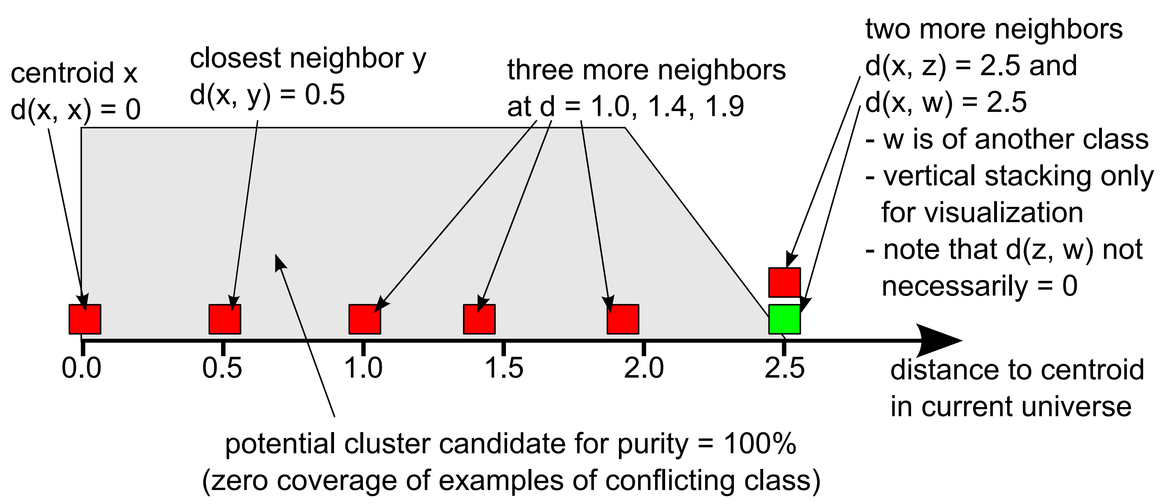
Images Descriptors on (i) color distribution, (ii) silhouettes or (iii) general meta information (image title, creation date, etc.)

3D Objects Descriptors (i. e. universes) encode properties of (i) shape, (ii) volume or (iii) connectivity of separated components

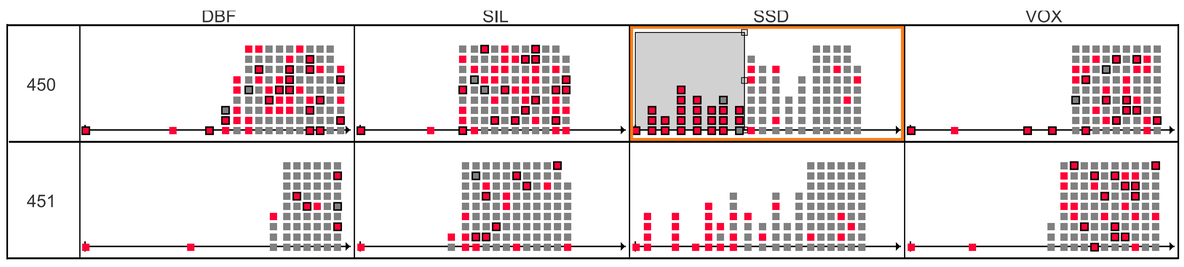
Other examples include the mining of web pages, the categorization of pieces of music or the analysis of proteins.

3 Supervised Approach – Neighborgrams [1]

- Neighborgram as data structure to reflect (distance-based) neighborhood around a centroid in a particular universe.
- Constructed for all objects of interest (e. g. minority class(es)) in all universes
- Derive cluster candidate from each Neighborgram, derive quality values (coverage for a given a cluster purity)
- Intuitive visualization enables interactive exploration

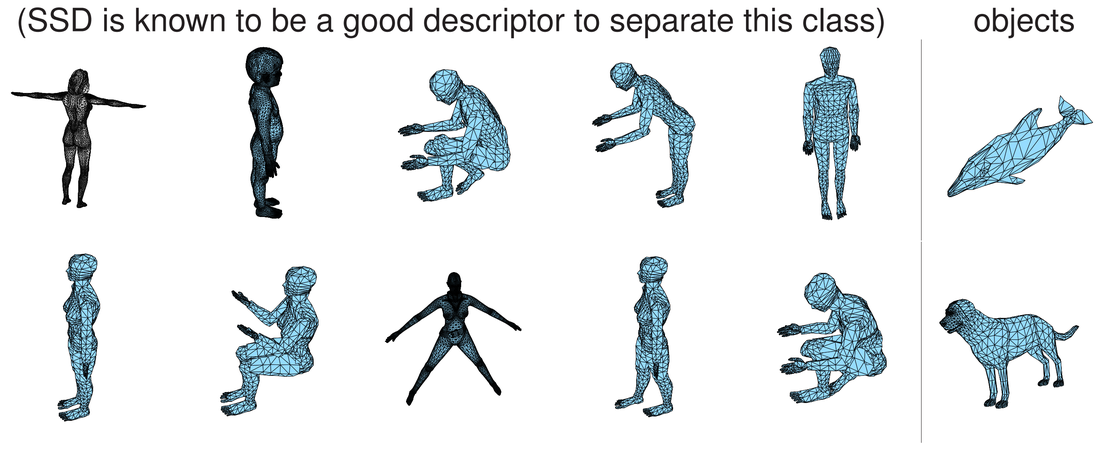


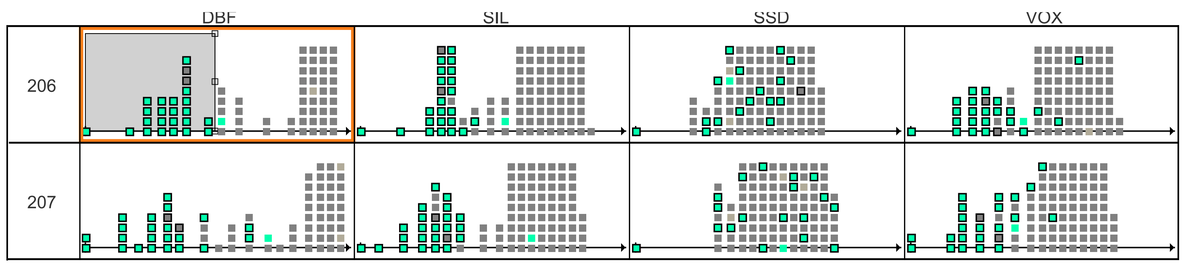
- Use ranking mechanism to automatically suggest "interesting" neighborgrams to user for further inspection (e. g. to assess cluster overlap across universes)
- Can also be run automatically, identifies universe-specific clusters and clusters spanning multiple universes
- Applied to data set of 3D objects (contributed by groups of Prof. Keim and Prof. Saupe), shown here are four universes: DBF and SIL (projection based), SSD (shape), and VOX (volume):



Cluster in SSD universe covers parts of the class human (SSD is known to be a good descriptor to separate this class)

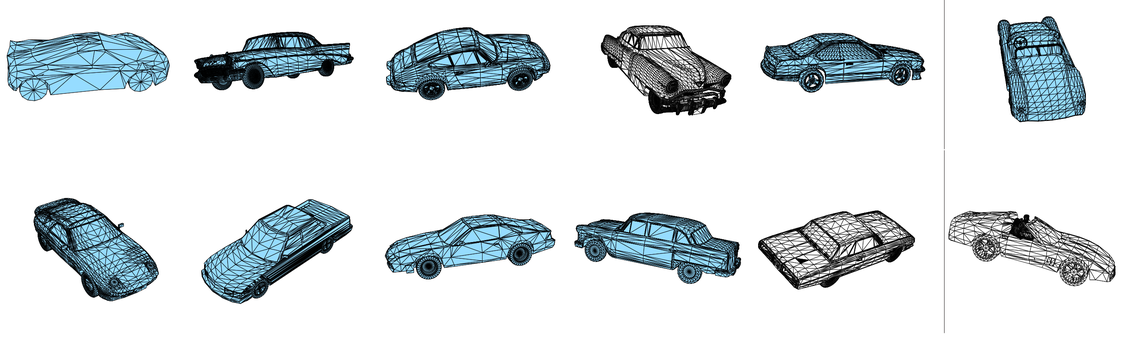
"borderline" objects





Cluster in DBF with strong overlap in SIL and VOX (cluster expansion limited by different car class (convertibles))

"borderline" objects



4 Unsupervised Approach - Extended Fuzzy c-Means [2]

- Objective function-based optimization problem, minimizing sum of distances $d_{ic}^{(u)}$ between objects i and cluster prototypes c, weighted by their degree of membership $v_{i,c}$
- Extension of classical objective function by additional term $z_{c,u}$ to reflect membership of cluster c to universe u

$$\sum_{c=1}^{C}\sum_{u=1}^{U} z_{c,u} \sum_{i=1}^{T} v_{i,c}\, d_{ic}^{(u)} \rightarrow \min! \text{ w.r.t. } \forall c : \sum_{u=1}^{U} z_{c,u} = 1,\ \forall i : \sum_{c=1}^{C} v_{i,c} = 1$$

- Similar idea applied to Possibilistic Clustering algorithm, added benefit: less affected by noise and bad choice of cluster count
- Methods applied to above mentioned 3D data set, majority of clusters correspond to pre-classification, most classes split into two or three clusters, some (very similar) classes were merged
- Future work: Address normalization problems, cluster validation, cluster count assessment

References

[1] Michael R. Berthold, Bernd Wiswedel, and David E. Patterson. Interactive exploration of fuzzy clusters using neighborgrams. *Fuzzy Sets and Systems*, 149(1):21–37, 2005.

[2] Bernd Wiswedel and Michael R. Berthold. Fuzzy clustering in parallel universes. *International Journal of Approximate Reasoning*, 45(3):439–454, 2007.