Correlation Clustering

Arthur Zimek
Ludwig-Maximilians-Universität München
Oettingenstraße 67, 80538 München, Germany
http://www.dbs.ifi.lmu.de/~zimek
zimek@dbs.ifi.lmu.de

ABSTRACT

This is a short summary of the author's thesis on "Correlation Clustering" (Ludwig-Maximilians-Universität München, Germany, 2008). The complete thesis is available at http://edoc.ub.uni-muenchen.de/8736/.

1. INTRODUCTION

While clustering in general is a rather dignified problem, mainly in about the last decade new approaches have been proposed to cope with new challenges provided by modern capabilities of automatic data generation and acquisition in more and more applications producing a vast amount of high dimensional data. These data need to be analyzed by data mining methods in order to gain the full potentials from the gathered information. However, high dimensional data pose different challenges for clustering algorithms that require specialized solutions. In particular, in high dimensional data traditional similarity measures as used in conventional clustering algorithms are often not meaningful. This problem and related phenomena triggered adaptations of clustering approaches to the nature of high dimensional data. Common approaches are known as subspace clustering, projected clustering, pattern-based clustering, or correlation clustering.[1] This area of research has been a highly active one in the recent years with a plethora of proposed algorithms but, in our opinion, lacking of a systematic problem analysis. Thus, a comparison of proposed algorithms is difficult both, theoretically and practically.

The family of axis-parallel subspace and projected clustering algorithms assumes that data objects belonging to the same cluster are near by each other in Euclidean space but allows to assess the corresponding distance of objects w.r.t. subsets of the attributes due to the problem of increasingly poor separation of near and far points in higher dimensional data and the problem of irrelevant attributes. Pattern-based approaches often disregard the assumption that a cluster consists of objects that are near by each other in the Euclidean space or some Euclidean subspace and, instead, aim at collecting objects following a similar behavioral pattern over a subset of attributes. These patterns relate to simple strict pairwise linear positive correlations among the considered attributes. *Correlation clustering* (sometimes also named *oriented clustering* or *generalized subspace/projected clustering*) generalizes this approach to arbitrary complex positive or negative correlations but often assumes, again, a certain density of the points in Euclidean space, too. These algorithms assume any cluster being located in an arbitrarily oriented affine subspace of the data space. Such clusters appear as hyperplanes of arbitrary dimensionality (linear manifolds) in the data space. However, no typical pattern in the data matrix does correspond to these models based on a spatial intuition. Thus, the approaches based on the intuitions typical for biclustering algorithms are in principle not suitable to tackle this more general problem, albeit some bicluster models can be regarded as very restricted special cases of correlation clustering models.

[1]Note, though, that the term "correlation clustering" has been used for a quite different problem within the machine learning community.

2. SUMMARY OF THE THESIS

The thesis "Correlation Clustering" aims at serving a twofold purpose. The first task was to systematically survey the very heterogenous field of clustering adaptations for high dimensional data. Second, new approaches have been contributed and integrated in the proposed systematics of algorithms and the reviewed problems typically encountered in the field.

Part I illustrates the background and motivation of the topic with the general context of data mining (Chapter 1) and some prominent application scenarios (Chapter 2).

Part II addresses the first main objective of the thesis in arranging the heterogeneous field of clustering algorithms for high dimensional data according to the typically addressed subproblems and approaches. To this end, Chapter 3 sketches the fundamental problem in clustering high dimensional data and characterizes three different classes of clustering algorithms. The subsequent chapters discuss these different classes of algorithms in more detail. First, axis-parallel clustering with the consuetudinary but questionable categorization in *projected* and *subspace clustering* is surveyed in Chapter 4. Pattern-based clustering algorithms are surveyed in Chapter 5. Here, the point was not to give an exhaustive overview on existing approaches but to point out the relationship and differences in comparison to subspace and correlation clustering. The latter class of algorithms has been surveyed in Chapter 6. Most of the approaches to this field have been developed by the author and constitute the second major contribution of the thesis. Chapter 7 concludes the systematic part with a discussion of the main problems and solutions. We have identified different aspects of the notorious "curse of dimensionality" that are addressed with different focus by different types of al-

gorithms. So far, no efficient all-round-algorithm has been proposed and it seems unlikely to get one. It is not even clear whether such a solution would be desirable. Another open question, as pointed out in this systematic part, is the evaluation of different approaches. Since any approach uses its own assumptions and heuristics (and often even defines the objective in a different way), a comprehensive and fair experimental evaluation of a reasonable set of representatives for the different classes of solutions is not only missing so far but seems also a very demanding and complex task.

Contributions to the category of correlation clustering algorithms are presented in Parts III–V. Part III collects adaptations of the density-based paradigm using PCA as a primitive to grasp correlated attributes and derive the corresponding arbitrarily oriented subspace. The first adaptation is the algorithm 4C (Chapter 8). A more robust, more efficient and more effective variant for flat correlation clustering is COPAC (Chapter 9). Hierarchical correlation clustering has been tackled by the approaches HiCO (Chapter 10) and ERiC (Chapter 11). For all correlation clustering algorithms based on PCA on a local selection of points, a framework to enhance the suitability of the selected set and the robustness of the applied PCA has been discussed in Chapter 12.

Nevertheless, as discussed in Part IV, there remain weak points of all these density-based approaches applying PCA on a local selection of representative points (see Chapter 13). They rely on the so called *locality assumption* which is in view of high dimensional data rather naïve, as has also been discussed in Part II. Thus, as a global approach to correlation clustering, the algorithm CASH has been proposed and discussed in Chapter 14.

None of the existing correlation clustering algorithms derives a quantitative and qualitative model for each correlation cluster. Such a model helps to gain the full practical potentials from correlation cluster analysis. Part V describes an original approach to derive quantitative information on the linear dependencies within correlation clusters. As discussed in Chapter 15, this step is not readily available for correlation clustering so far. The concepts for deriving quantitative and qualitative correlation clustering models described in Chapter 16 are independent of the clustering approach and can thus be applied as a post-processing step to any correlation clustering algorithm. The broad experimental evaluation demonstrates the beneficial impact of the proposed method on several applications of significant practical importance. It has been exemplified how the method can be used in conjunction with a suitable clustering algorithm to gain valuable and important knowledge about complex relationships in real-world data. Furthermore, as sample applications of the approach, Chapter 17 sketches how these quantitative models can be used to predict the probability distribution that an object is created by these models, and Chapter 18 describes an adaptation of the approach to the outlier detection problem.

The final Part VI summarizes the contributions and results of the thesis (Chapter 19), and points out some open questions and possible directions for future work (Chapter 20).

3. PUBLICATIONS OF PARTIAL ASPECTS

The results of Part II were presented as tutorial at ICDM '07, PAKDD '08, KDD '08, and VLDB '08 [11], and have been published in an updated version in [12]. The algorithms developed in Parts III and IV have also been published in [9; 6; 7; 5; 10; 3; 2]. The major content of Part V has been presented in [4]. The implementations of all algorithms are available in the framework ELKI [8; 1]

4. ACKNOWLEDGEMENTS

I wish to express my sincerest gratitude to my supervisor, Hans-Peter Kriegel, and to Thomas Seidl, who served as secondary supervisor. Furthermore, I acknowledge all the help and contributions I received from my coauthors and colleagues.

5. REFERENCES

[1] E. Achtert, T. Bernecker, H.-P. Kriegel, E. Schubert, and A. Zimek. ELKI in time: ELKI 0.2 for the performance evaluation of distance measures for time series. In *Proc. SSTD*, 2009.

[2] E. Achtert, C. Böhm, J. David, P. Kröger, and A. Zimek. Global correlation clustering based on the hough transform. *Stat. Anal. Data Min.*, 1(3):111–127, 2008.

[3] E. Achtert, C. Böhm, J. David, P. Kröger, and A. Zimek. Robust clustering in arbitrarily oriented subspaces. In *Proc. SDM*, 2008.

[4] E. Achtert, C. Böhm, H.-P. Kriegel, P. Kröger, and A. Zimek. Deriving quantitative models for correlation clusters. In *Proc. KDD*, 2006.

[5] E. Achtert, C. Böhm, H.-P. Kriegel, P. Kröger, and A. Zimek. On exploring complex relationships of correlation clusters. In *Proc. SSDBM*, 2007.

[6] E. Achtert, C. Böhm, H.-P. Kriegel, P. Kröger, and A. Zimek. Robust, complete, and efficient correlation clustering. In *Proc. SDM*, 2007.

[7] E. Achtert, C. Böhm, P. Kröger, and A. Zimek. Mining hierarchies of correlation clusters. In *Proc. SSDBM*, 2006.

[8] E. Achtert, H.-P. Kriegel, and A. Zimek. ELKI: a software system for evaluation of subspace clustering algorithms. In *Proc. SSDBM*, 2008.

[9] C. Böhm, K. Kailing, P. Kröger, and A. Zimek. Computing clusters of correlation connected objects. In *Proc. SIGMOD*, 2004.

[10] H.-P. Kriegel, P. Kröger, E. Schubert, and A. Zimek. A general framework for increasing the robustness of PCA-based correlation clustering algorithms. In *Proc. SSDBM*, 2008.

[11] H.-P. Kriegel, P. Kröger, and A. Zimek. Detecting clusters in moderate-to-high dimensional data: subspace clustering, pattern-based clustering, and correlation clustering. *PVLDB*, 1(2):1528–1529, 2008.

[12] H.-P. Kriegel, P. Kröger, and A. Zimek. Clustering high dimensional data: A survey on subspace clustering, pattern-based clustering, and correlation clustering. *ACM TKDD*, 3(1):1–58, 2009.