Detecting Clusters in Moderate-to-High Dimensional Data: Subspace Clustering, Pattern-based Clustering, and Correlation Clustering

Hans-Peter Kriegel Peer Kröger Arthur Zimek
Ludwig-Maximilians-Universität München
Oettingenstr. 67, 80538 München, Germany
http://www.dbs.ifi.lmu.de
{kriegel,kroegerp,zimek}@dbs.ifi.lmu.de

ABSTRACT

As a prolific research area in data mining, subspace clustering and related problems induced a vast amount of proposed solutions. However, many publications compare a new proposition – if at all – with one or two competitors or even with a so called "naïve" *ad hoc* solution but fail to clarify the exact problem definition. As a consequence, even if two solutions are thoroughly compared experimentally, it will often remain unclear whether both solutions tackle the same problem or, if they do, whether they agree in certain tacit assumptions and how such assumptions may influence the outcome of an algorithm. In this tutorial, we try to clarify (i) the different problem definitions related to subspace clustering in general, (ii) the specific difficulties encountered in this field of research, (iii) the varying assumptions, heuristics, and intuitions forming the basis of different approaches, and (iv) how several prominent solutions essentially tackle different problems.

1. INTRODUCTION

Clustering aims at dividing data sets into subsets (clusters), maximizing intra-cluster-similarity while minimizing inter-cluster-similarity of objects. 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 of high dimensional data. This new family of algorithms is not yet backed by a systematic problem analysis. The measure of similarity, however, is probably the most important difference among different algorithms. The used measure of similarity does not only influence the outcome of an algorithm but is also of relevance for judging the quality of a resulting clustering. Thus, a comparison of proposed algorithms is difficult both, theoretically and practically.

Recently, some surveys have already given overviews on some approaches. For example, in [14], some basic problems are illustrated and some approaches are sketched. However, there is no clear distinction between different subproblems (axis-parallel or arbitrarily oriented) and the corresponding algorithms are discussed without pointing out the underlying differences in the corresponding problem definitions. In [13], the focus is on pattern-based clustering approaches and the specialized application domain of microarray data.

Here, we would like to give a more systematic approach to the problem and on the different tasks and subproblems (axis-parallel, pattern-based, correlation clustering). Therefore, we will also survey the related heuristics used by different approaches. Our systematic view is not based on the application scenarios but on the intrinsic methodological differences of the various families of approaches based on different spatial intuitions. Thus, we will also try to integrate the inherently different point of view of pattern-based approaches into the intuition of patterns in the data space.

Longer versions of this tutorial were presented at ICDM 2007, PAKDD 2008, and KDD 2008.

PVLDB '08, August 23-28, 2008, Auckland, New Zealand
Copyright 2008 VLDB Endowment, ACM 978-1-60558-306-8/08/08

2. BASIC PROBLEMS

High dimensional data confronts cluster analysis with several problems. A bundle of problems is commonly addressed as the "curse of dimensionality". Aspects of this "curse" most relevant to the clustering problem are: (i) Any optimization problem becomes increasingly difficult with an increasing number of variables (attributes) [7]. (ii) The relative distance of the farthest point and the nearest point converges to 0 with increasing data dimensionality [8, 11], i.e., the discrimination between the nearest and the farthest neighbor becomes rather poor in high dimensional data spaces. (iii) Automated data acquisition in many application domains leads to the collection of as many features as possible. Many of these features may eventually provide useful insights but for the task at hand in many problems there exist many irrelevant attributes in a data set. Since groups of data are defined by some of the attributes only, the remaining irrelevant attributes ("noise") may heavily interfere with the efforts to find these groups. (iv) Similarly, in a data set containing many attributes, some attributes will most likely exhibit correlations among each other (in varying complexity).

Many approaches try to alleviate the "curse of dimensionality" by applying feature selection or dimensionality reduction methods prior to cluster analysis. However, the second main challenge for cluster analysis of high dimensional data

is the possibility and even high probability that different subsets or combinations of attributes may be relevant for different clusters. Thus, a global feature selection or dimensionality reduction method cannot be applied. Rather, it becomes an intrinsic problem of the clustering approach to find the relevant subspaces and to find clusters in these relevant subspaces. Furthermore, although correlation among attributes often is the basis for a dimension reduction, for many application domains it is a main part of the interesting information what correlations exist among which attributes for which subsets of objects. As a consequence of this second challenge, the first challenge (i.e., the "curse of dimensionality") generally cannot be alleviated for clustering high dimensional data by global feature selection or global dimensionality reduction.

3. COVERED MODELS AND REPRESENTATIVE APPROACHES

Subspace clustering techniques can be divided into three main families. In view of the challenges sketched above, any arbitrarily oriented subspace may be interesting for a subspace clustering approach. The most general techniques ("(arbitrarily) oriented clustering", "correlation clustering") tackle this infinite search space. Example algorithms are described in [5, 9, 1], the general model for this family of approaches is described in [3]. Yet most of the research in this field assumes the search space to be restricted to axis-parallel subspaces. Since the search space of all possible axis-parallel subspaces of a d-dimensional data space is still in $O(2^d)$, different search strategies and heuristics are implemented. Axis-parallel approaches mainly split into "subspace clustering" and "projected clustering". Examples here are [6, 4, 12, 2]. In between these two main fields a group of approaches is known as "pattern-based clustering" (also: "biclustering" or "co-clustering"). For these approaches, the search space is not necessarily restricted to axis-parallel subspaces but on the other hand does not contain all arbitrarily oriented subspaces. The restrictions on the search space differ substantially between different approaches in this group. Prominent algorithms are described in [10, 16, 15].

The family of axis-parallel subspace and projected clustering algorithms assumes that data objects belonging to the same cluster are close to each other 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 close to 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 usually relate to simple positive correlations among the considered attributes. Correlation clustering approaches generalize this approach to arbitrarily complex positive or negative correlations but often (except for [1]) assume, again, a certain density of the points in Euclidean space, too.

4. CONCLUSION

The aim of the concrete task of data analysis influences the choice of the clustering algorithm and obviously also the interpretation of the results of the clustering process. The appropriate choice of a clustering approach adequate to the problem at hand should be based on knowledge of the basic principles the particular clustering approach is based upon. Similarly, the interpretation of clustering results should be guided by the knowledge of the kinds of patterns a particular algorithm can or cannot find. This tutorial aims mainly at characterizing the different underlying assumptions and models for these different yet related families of clustering algorithms and at supporting such decisions and interpretations by a systematic overview on the different kinds of algorithms specialized to different problems known to occur in high dimensional data.

5. REFERENCES

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

[2] E. Achtert, C. Böhm, H.-P. Kriegel, P. Kröger, I. Müller-Gorman, and A. Zimek. Detection and visualization of subspace cluster hierarchies. In *Proc. DASFAA*, 2007.

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

[4] C. C. Aggarwal, C. M. Procopiuc, J. L. Wolf, P. S. Yu, and J. S. Park. Fast algorithms for projected clustering. In *Proc. SIGMOD*, 1999.

[5] C. C. Aggarwal and P. S. Yu. Finding generalized projected clusters in high dimensional space. In *Proc. SIGMOD*, 2000.

[6] R. Agrawal, J. Gehrke, D. Gunopulos, and P. Raghavan. Automatic subspace clustering of high dimensional data for data mining applications. In *Proc. SIGMOD*, 1998.

[7] R. Bellman. *Adaptive Controll Processes. A Guided Tour*. Princeton University Press, 1961.

[8] K. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft. When is "nearest neighbor" meaningful? In *Proc. ICDT*, 1999.

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

[10] Y. Cheng and G. M. Church. Biclustering of expression data. In *Proc. ISMB*, 2000.

[11] A. Hinneburg, C. C. Aggarwal, and D. A. Keim. What is the nearest neighbor in high dimensional spaces? In *Proc. VLDB*, 2000.

[12] H.-P. Kriegel, P. Kröger, M. Renz, and S. Wurst. A generic framework for efficient subspace clustering of high-dimensional data. In *Proc. ICDM*, 2005.

[13] S. C. Madeira and A. L. Oliveira. Biclustering algorithms for biological data analysis: A survey. *IEEE TCBB*, 1(1):24–45, 2004.

[14] L. Parsons, E. Haque, and H. Liu. Subspace clustering for high dimensional data: A review. *SIGKDD Explorations*, 6(1):90–105, 2004.

[15] J. Pei, X. Zhang, M. Cho, H. Wang, and P. S. Yu. MaPle: A fast algorithm for maximal pattern-based clustering. In *Proc. ICDM*, 2003.

[16] H. Wang, W. Wang, J. Yang, and P. S. Yu. Clustering by pattern similarity in large data sets. In *Proc. SIGMOD*, 2002.