Mining Subspace Clusters: Enhanced Models, Efficient Algorithms and an Objective Evaluation Study

Emmanuel Müller

supervised by Prof. Thomas Seidl

Data Management and Data Exploration Group
RWTH Aachen University, Germany
{mueller, seidl}@cs.rwth-aachen.de

ABSTRACT

In the knowledge discovery process, clustering is an established technique for grouping objects based on mutual similarity. However, in today's applications for each object very many attributes are provided in large and high dimensional databases. As multiple concepts described by different attributes are mixed in the same data set, clusters are hidden in subspace projections and do not appear in all dimensions. *Subspace clustering* aims at detecting such clusters in any projection of the database.

This work presents an overview of my dissertation on subspace clustering models, efficient processing schemes for these models and an objective evaluation study on subspace clustering techniques. This work highlights several open challenges and our research work with which we tackled these challenges. Furthermore, as a general contribution to the community, the benefits of our evaluation study and our open source evaluation framework are described. Both provide an important basis for future research and ensure comparability and repeatability of experiment results.

1. INTRODUCTION

In the knowledge discovering process, clustering aims at detecting groups of similar objects while separating dissimilar ones. Traditional clustering approaches compute a partition of the data, grouping each object in at most one cluster or detecting it as noise. However, it is not always the case that an object is part of only one cluster. Multiple meaningful groupings might exist for each object. The detection of such multiple clusters describing different views on each object is still an open challenge in recent applications.

In today's applications, data is collected for multiple analysis tasks. In most cases, databases contain objects specified by very many attributes. As one does not know the hidden structure of the data, one mixes up different measurements in one high dimensional database. Thus, each object can participate in various groupings reflected in different subsets of the attributes. For example, in customer segmentation, objects are customers described by multiple attributes specifying their profile. A customer might be grouped by the attributes "average traveling frequency" and "income" with other "globetrotters" having high values in both of these attributes. The same customer might be a "healthy oldie" which could be specified by a high "age" and low "blood pressure" (cf. Fig. 1). We observe for each customer multiple possible behaviors which should be detected as clusters. Thus, clusters may overlap in their clustered objects, i.e. each object may be represented in multiple clusters. Furthermore, each behavior of a customer is described by specific attributes. Thus, meaningful clusters appear only in these specific subspace projections of the data. While the attribute "blood pressure" is useful for the distinction of health status, the attribute "traveling frequency" might be irrelevant for health related groups of customers.

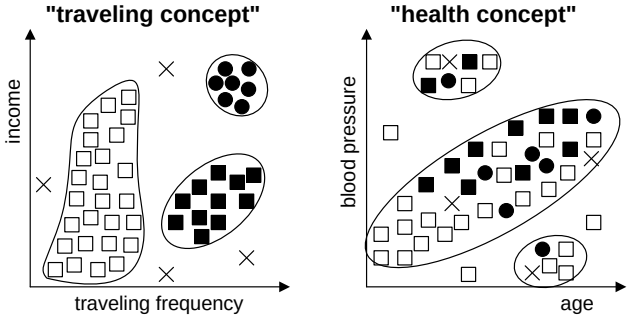

Figure 1: Example of different subspace clusters

We generalize these observations as they are not only applicable to customer segmentation. In other applications, objects might be sensor nodes represented by multiple sensor measurements, or objects might be genes described by their expression level under multiple conditions. For each of these application scenarios, objects are described by very many attributes. For such high dimensional data, all objects seem to be unique in full space as distances grow alike due to the so called "curse of dimensionality" [8]. However, a common observation is that each of the objects might be part of different groups in different subsets of attributes. In

VLDB 2010 PhD Workshop, September 13, 2010, Singapore

general, we call this an object that is part of multiple *concepts* detected by clusters in different subspace projections. All of these groupings are valid characterizations of the same objects by using different attributes. Thus, for the general case of high dimensional data, a subspace cluster can be detected by considering a subset of the dimensions.

Formally, a subspace cluster is a group of objects considering a set of relevant attributes. The main characterization of a subspace cluster is given by its set of relevant dimensions in which the objects are grouped together.

DEFINITION 1. ***Subspace Cluster***
Given a database DB which describes each object $o \in DB$ with dimensions DIM. A subspace cluster $C = (O, S)$ is defined by a set of objects $O \subseteq DB$ and a subset of dimensions $S \subseteq DIM$.

General research questions in subspace clustering consider novel clustering models, their efficient computation and fair and objective evaluation of the resulting output. All three of these topics are discussed in the following.

2. RELATED WORK

Different clustering paradigms have been proposed in the past decades for clustering and the young research area of subspace clustering. *Traditional clustering* approaches, aim at the detection of clustered objects using all attributes in the full data space. However, independent of the underlying clustering model, full space clustering approaches do not scale to high dimensional data spaces covering multiple different concepts. As clusters do not appear across all attributes, they are hidden by irrelevant attributes. Dimensionality reduction like PCA aims at discarding irrelevant, noisy dimensions [13]. However, in many practical applications no globally irrelevant dimensions exist.

Recent research for clustering in high dimensional data has introduced a number of different approaches summarized in [24, 16]. The underlying mining task was named by the pioneers in this field *subspace clustering* [2] or *projected clustering* [1]. Their common goal is to detect clusters in arbitrary subspace projections of the data. Each cluster is associated with a set of relevant dimensions in which this pattern has been discovered. Differences of clustering approaches in subspace projections have been shown in our recent evaluation study [23]. *Projected clustering* techniques detect disjoint subsets of objects [1, 17, 25, 29]. Thus, projected clustering misses to detect multiple concepts per object as one aims at a partitioning of the objects. In contrast, *subspace clustering* allows objects to be part of multiple clusters in arbitrary subspaces. However, due to the exponential number of possible subspaces, it results in a huge number of redundant clusters [2, 26, 14, 15].

In general, one might observe that most of the proposed techniques in the literature have focused on extending traditional cluster definitions to subspace projections. Thus, the state-of-the-art methods have missed to address some specific challenges in subspace clustering. Especially, the detection of few but relevant concepts in subspace projections has not yet been addressed in subspace clustering. And thus, redundancy as a major challenge has not been addressed for almost a decade of research in subspace clustering. Let us summarize the most important challenges in subspace clustering in the following section before addressing our techniques to tackle these challenges.

3. SUBSPACE CHALLENGES

Abstracting from the mentioned application scenarios there are several open challenges in the area of subspace clustering. Let us summarize these challenges to derive an overview of requirements for our novel subspace clustering techniques. The following sections then focus on these efficient and accurate techniques tackling each of these challenges.

CHALLENGE 1. ***Adaptive cluster definition***

The first challenge is derived out of the cluster definition itself. While traditional clustering methods consider only one space (full data space) they provide one global cluster definition. Each set of objects which fulfills this definition is a cluster. In contrast, subspace clustering searches for clusters (O, S) in arbitrary subspaces S. Using one cluster definition for all of these subspaces might miss some meaningful clusters. A subspace cluster definition should *adapt* to the considered subspace. Traditional subspace clustering approaches only restrict their (dis-)similarity measure to dimensions in S (e.g. Euclidian Distance $dist_S(o,p) = \sqrt{\sum_{i \in S}(o_i - p_i)^2}$ is restricted to dimensions in S). However, such distances are incomparable over different subspaces (e.g. for $T \subseteq S : \; dist_T(o,p) \leq dist_S(o,p)$). In general, traditional subspace clustering approaches are biased with respect to the dimensionality of the considered subspaces [3]. As main property of each subspace, the dimensionality has major influence on the data distribution, and thus, it should be utilized to develop an unbiased subspace cluster definition. As distances grow with increasing dimensionality, objects are expected to be dense in low dimensional subspaces while scattered in high dimensional spaces. To detect meaningful clusters in any dimensionality, cluster definition should adapt w.r.t. to this phenomenon.

CHALLENGE 2. ***Detection of multiple concepts***

A subspace clustering as the final set of subspace clusters should allow the detection of multiple clusters for each object. Detecting groups of objects in arbitrary subspaces provides a set of attributes as reasons for each cluster. Each subspace represents one of the multiple hidden concepts in a high dimensional database. As observed in several application scenarios, each object might be part of multiple clusters in different subspaces. Assigning each object to at most one cluster would restrict subspace clustering to only one concept and result in missing clusters in other subspaces. A subspace clustering should allow the detection of multiple concepts. We consider overlapping of subspace clusters as a major requirement. Furthermore, in subspace clustering one should actively search for different concepts hidden in the data. The general aim is to detect multiple subspace clusters for all objects each of them representing a different view on the data.

CHALLENGE 3. ***Redundancy of subspace clusters***

In contrast to the benefits of detecting multiple concepts, one has to cope with enormous amount of possible clusters in arbitrary subspaces. In an ideal case, all multiple concepts provide additional knowledge for the overall result set. However, this is not true for redundant subspace clusters. Each hidden subspace cluster (O, S) results in a tremendous amount of redundant subspace clusters $(O, T) \; \forall \; T \subset S$ in

all its lower dimensional projections providing no additional knowledge. Removal of such redundancy is an important challenge for subspace clustering to reduce the result to few but relevant subspace clusters. There are several redundancy issues to solve, ranging from simple projections of subspace clusters inducing redundant results up to optimization problems of overall result set.

CHALLENGE 4. ***Computational complexity***

While the previous challenges address the quality of subspace clustering results, one has also to consider an efficient computation. In contrast to traditional clustering approaches which have to show scalability with respect to increasing number of objects in the database, subspace clustering techniques additionally have to scale with the number of given attributes per object. Using the density-based paradigm one requires $O(|DB|^2)$ for clustering one fixed subspace, while there are $2^{|DIM|} - 1$ many possible subspaces to investigate. Reducing both of these computational costs by pruning cluster candidates is essential for a good runtime performance, and thus, also for the applicability of subspace clustering to large and high dimensional databases. We show that costly database access like in traditional approaches and the exponential search space of arbitrary subspaces pose challenges for an efficient computation. Especially, for our enhanced optimization models we have proven that mining the most relevant subspace clusters is NP-hard [19]. This poses novel challenges for the development of approximative efficient algorithms.

CHALLENGE 5. ***Exploration and evaluation***

As a natural property for clustering (unsupervised learning), no knowledge is given about the hidden structure of the data. This poses a major challenge to evaluation of subspace clustering results. One possible but quite subjective way of evaluation is visual exploration of results by domain experts. However, for the young research area of subspace clustering exploration tools are not available but strongly desired. A second more objective way of evaluation is the use of labeled data assuming that the given class labels represent the hidden cluster structure. While labeled data are widely used for cluster evaluation, there exists no systematic evaluation of subspace clustering techniques. Especially, due to missing standardized evaluation measures and a missing comparability of different implementations the comparison of subspace clustering techniques is highly challenging. Overall, the empirical results in most of the scientific publications on subspace clustering do not provide any objective and systematic comparison. Different paradigms coexist in the literature without any empirical evaluation of their clustering qualities.

4. NOVEL CLUSTERING MODELS

We observe multiple open challenges to be tackled in this young research area. In our research we focus on efficient and accurate subspace clustering models based on the density-based paradigm. In the following, an overview of this research work tackling each of the mentioned challenges is given. As an overview only the basic ideas are provided in this work, for detailed discussions and evaluation results please refer to the original publications.

4.1 Adaptive density-based subspace clustering

As first contribution, we provide a subspace cluster definition which adapts to the considered subspace. Focusing on the density-based paradigm proposed by DBSCAN for the full space clustering [9], we develop an adaptive density which is aware of the decreasing expected density in higher dimensional spaces.

The traditional density-based clustering defines clusters as dense regions seperated by sparse areas. Density is measured by simply counting the objects in a fixed ε-neighborhood: $den_S(o) = |\{p \in DB \mid dist_S(o,p) \leq \varepsilon\}|$ Besides the restriction to the subspace S this is the original definition as provided by DBSCAN [9]. However, for increasing dimensionality $|S|$ subspace clustering models have to cope with increasing distances and the decreasing densities. In general, our density models automatically adapt to the expected density in each subspace and represent density properties of the hidden clusters better than fixed density definitions.

As basic enhancement we propose to adapt the density threshold in the density-based cluster criterion. Intuitively, clusters are defined as dense objects [9], that have to exceed a certain density threshold $den_S(o) \geq MinPts$. Traditionally, this threshold is fixed, but as density drops for increasing dimensionality we define a monotonically decreasing threshold function. Our threshold adapts to the expected density distribution caused by the dimensionality $|S|$ of the considered subspace.

DEFINITION 2. ***Unbiased density***
An object $o \in DB$ is called denser than expected in the subspace S iff:

$$den_S(o) \geq expected\_density(S)$$

Our threshold function adapts the cluster definition such that objects have to be denser than expected in the considered subspace. As we have shown in previous work [3], this unbiased density enhances the quality of subspace clustering results. Furthermore, we have extended this density definition given only for continuous valued attributes to cope with heterogeneous subspaces [22]. We developed a unified density for both continuous and categorical attributes. We provide a thorough comparison between frequent itemset and subspace clustering as different mining paradigms. We derive common properties for frequency on categorical data and density on continuous data and unify these two mining paradigms for our heterogeneous subspace clustering model. This is essential for real world data where typically attributes have various different types. As further enhancements we develop an adaptive density measure.

$$adaptive\_density(o,S) = |\{p \in DB \mid dist(o,p) \leq \varepsilon(S)\}|$$

While the previous approaches adapt the density threshold and keep measuring the density in a fixed neighborhood, our final approach defines a novel density measure. It increases the neighborhood by a statistically motivated function $\varepsilon(S)$ according to the dimensionality of the subspace [19, 27]. With this enhanced density measure we ensure meaningful density values even for subspaces where traditional fixed neighborhoods tend to become empty. Overall, our research introduces novel techniques for density measures designed specifically for subspace clustering.

4.2 All and only relevant subspace clusters

As second contribution we provide a novel clustering definition for the final set of resulting subspace clusters. The major enhancement is due to the exclusion of redundant clusters. As we include only few but relevant subspace cluster we improve the clustering quality and in addition gain efficiency improvements for our algorithmic solutions. Existing projected and subspace clustering algorithms do not address redundancy handling adequately. Projected clustering simply forces results to be non-redundant by assigning each object to a single cluster at the cost of missing overlapping clusters. Subspace clustering algorithms, in contrast, either use no or a mere local approach to check the redundancy.

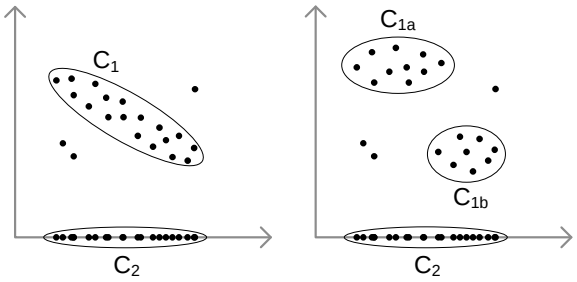

Figure 2: Local and global redundant clusters

As first approach in this area we develop a local redundancy definition excluding lower dimensional projections of subspace clusters (cf. Fig. 2). If for example a cluster (C_2) covers nearly the same objects than a higher dimensional cluster (C_1), then (C_2) is redundant and not output as a result. Both clusters would have been valid results if one considers only the cluster definition without redundancy handling. However, excluding redundant clusters ensures small and high quality clustering results. Our local redundancy definition as proposed in [6], excludes lower dimensional projections that do not contribute enough novel information controlled by a parameter R.

DEFINITION 3. ***Non-Redundant Subspace Clustering*** *A subspace cluster (O, S) is non-redundant iff:*

$$\nexists (O', S') \text{ with } O' \subseteq O \wedge S' \supset S \wedge |O'| \geq R \cdot |O|$$

This simple redundancy handling has shown improvements of both clustering quality and runtime performance due to in-process removal of redundant clusters [6]. However it is limited to local redundancy. As shown in our example, in both subfigures the cluster C_2 is redundant because it is induced by the other clusters C_1, resp. C_{1a}, C_{1b}. A local approach could identify the redundancy in the left figure. Cluster C_2 is redundant, as it covers C_1 and only a few additional objects. In the right figure, the fraction of points shared by C_{1a} and C_2 as well as by C_{1b} and C_2 is small, and the cluster C_2 is misleadingly classified as non-redundant. This mistake is the result of the local view on redundancy, i.e. for each check only a pairwise comparison of clusters is performed

In our second redundancy definition we overcome the drawbacks of pairwise comparison and ensure redundancy-free clustering by a global optimization. This enables the redundancy model to exclude also redundant clusters as depicted in Fig. 2. A cluster is only included if it contributes novel knowledge w.r.t. all other clusters in the clustering result [19]. In contrast to our first approach, this optimization yields only quality improvements while we have proven that it is an NP-hard problem. This poses new challenges for the development of efficient approximative solutions (cf. Challenge 4). Based on this optimization we further enhance our clustering model by comparing each cluster only with all other clusters in similar subspaces. Very dissimilar subspaces (orthogonal subspaces) provide novel knowledge as different concepts might be hidden in these subspaces (cf. Challenge 2). Thus, our orthogonal subspace clustering model [12], actively includes novel knowledge of orthogonal concepts into the final clustering result. In ongoing work [11, 10], we extend this to detection of alternative clusters in subspaces projections. Overall, all of our non-redundant subspace clustering techniques enable a high quality selection of relevant subspace cluster and thus significantly reduce the result sizes of subspace clustering.

5. EFFICIENT ALGORITHMS

As third contribution we develop several novel solutions for efficient computation of our enhanced subspace clustering models. In general, these solutions tackle the high cost for database access, prune the exponential search space of arbitrary subspace projections and propose efficient solutions for optimization of clustering result. Overall, these solutions scale well with increasing number of dimensions, which has shown to be the most challenging task for efficient processing. Evaluated on benchmark data sets these solutions show practical runtimes for the computation of high quality clustering results.

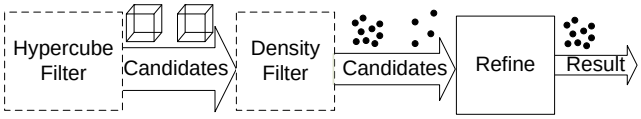

Figure 3: Efficient filter and refinement architecture

As basic solution to tackle the high cost of database access we propose a filter and refinement architecture for subspace clustering (cf. Fig. 3). This basic idea uses a grid approximation as filter and a density-based clustering as a refinement step [5]. As filter steps can be performed without costly database access, our architecture ensures scalable runtimes up to large and high dimensional databases. It is used in all of our further developments to ensure efficient access to arbitrary subspace regions. Using grid approximations our techniques can efficiently check whether a subspace region contains a potential cluster or not. Thus pruning of irrelevant regions can be performed without costly database access.

Incorporating our redundancy removal into an efficient processing we developed a pruning of redundant subspace regions in a depth-first processing [6]. Using this technique one can exclude large parts of the exponential search space as they contain only redundant clusters. This pruning can be realized to perform an in-process removal of redundant clusters without creating tremendous amount of redundant subspace cluster candidates. For the more enhanced subspace clustering models using an optimization of the clustering result we have to cope with an NP-hard problem, as mentioned before. We develop a relaxation of the proposed model [19].

This relaxation can be computed by a greedy processing showing both efficient computation and high quality clustering results. As our relevant subspace clustering excludes most of the exponential search space, in our ongoing work we develop a jump processing for subspace clustering. This novel processing schema overcomes the efficiency problems of both traditional breadth-first processing as well our own depth-first processing methods. A key requirement for such a jump is a high quality estimation of density in arbitrary subspace regions. We develop an efficient but also high quality estimation [20]. Using only knowledge from two dimensional histograms our method can estimate density in subspace regions without further database access. This extends our basic filter and refinement architecture by a second filter step with even less computation cost.

6. OBJECTIVE EVALUATION

As general contribution to the research community we have performed a systematic evaluation of a broad set of subspace clustering techniques. Furthermore, our novel framework for evaluation and exploration of clustering in subspace projections provides the means for a comparable and especially a repeatable evaluation of the main paradigms.

6.1 Evaluation Study

In our evaluation study [23], we provide a thorough characterization of the main properties of subspace clustering paradigms and their instantiations in different approaches. We provide an overview of three major paradigms (cell-based, density-based and clustering oriented). We highlighted important properties for each of these paradigms and compare them in extensive evaluations. In a systematic evaluation we used several quality measures and provide results for a broad range of synthetic and real world data.

With this study, we provide the first comparison of different paradigm properties in a thorough evaluation. We could show that traditional methods for density-based subspace clustering do not scale to very high dimensional data, while the clustering oriented approaches are affected by noisy data resulting in low clustering quality. A recent cell-based approach [29], outperforms in most cases the competitors in both efficiency and clustering quality. Surprisingly, the basic approach PROCLUS [1], in the clustering oriented paradigm, performs very well in our comparison. In contrast, the basic approaches CLIQUE and SUBCLU of the other two paradigms showed major drawback induced by the tremendously large result set. Recent approaches of these paradigms enhanced the quality and efficiency, however, could reach top results only in few cases. Summing up, we show that computing only a small set of relevant clusters like some projected clustering approaches and pruning most of the redundant subspace clusters as proposed by our techniques achieves best results.

In general, our evaluation constitutes an important basis for subspace clustering research as one can compare a broad set of competing techniques but also compare the used evaluation measure. Using this study one can derive several novel challenges not tackled by other subspace clustering approaches. Furthermore, we observe ongoing publications in this area for which our study gives a baseline for future evaluations. Our proposed baseline includes multiple aspects for a fair comparison not only in evaluation studies: First, a common open source framework with baseline implementations for a fair comparison of different algorithms. Second, a broad set of evaluation measures for clustering quality comparison. Third, a baseline of evaluation results for both real world and synthetic data sets with given parameter settings for repeatability. All of this can be downloaded from our website[1] for further research, comparison or repeatability. In the following we describe the underlying open source framework in more details. It is a major contribution for the community as it enables repeatable evaluation for future publications. Furthermore, this common framework ensures applicability of our research in various domains by providing evaluation and exploration of subspace clustering results.

6.2 Open Source Framework

With OpenSubspace we provide an open source framework for the emerging research area of clustering in subspace projections [18, 7]. The aim of our framework is to establish a basis for comparable and repeatable experiments and thorough evaluations in the area of clustering on high dimensional data. OpenSubspace is designed as the basis for comparative studies on the advantages and disadvantages of different subspace/projected clustering algorithms. Providing OpenSubspace as open source, our framework can be used by researchers and educators to understand, compare, and extend subspace and projected clustering algorithms. The integrated state-of-the-art performance measures and visualization techniques are first steps for a thorough evaluation of algorithms in this field of data mining.

As major benefit OpenSubspace ensures repeatability of results in scientific publications. Based on a common framework all implementations are available and can be used by any researcher to compare against previous techniques. As integrated in the well established WEKA framework [28], our OpenSubspace tool is already used by several international scientists for their research work. Especially for our research it is widely used in evaluations. All experiments in our subspace clustering publications are based on this framework such that we can ensure repeatability of our results. Furthermore, we use this framework in lab courses to provide students an easy way of rapid prototyping as well as in lectures to illustrate the effects of subspace clustering algorithms on toy databases.

Overall, the OpenSubspace framework can be seen as the natural basis for our future research in this area. We are currently developing evaluation measures that meet the requirements for a global quality rating of subspace clustering results. Evaluations with the given measurements show that none of the measurements can provide an overall rating of quality. Some measurements give contradicting quality ratings on some data sets. Such effects show us that further research should be done in this area.

Visualization techniques give an overall impression on the groupings detected by the algorithms [4, 21]. Further research of meaningful and intuitive visualization is clearly necessary for subspace clustering. The open source framework might encourage also researches in Visual Analytics to develop more meaningful visualization and exploration techniques.

For an overall evaluation framework OpenSubspace provides algorithm and evaluation implementations. However, further work has to be done to collect a bigger test set of

[1] http://dme.rwth-aachen.de/OpenSubspace/

high dimensional data sets. On such a benchmarking set one could collect best parameter settings and best quality results for various algorithms as example clusters on these data sets. The aim of an overall evaluation framework with benchmarking data will then lead to a more mature subspace clustering research field in which one can easily judge the quality of novel algorithms by comparing it with approved results of competing approaches.

7. CONCLUSION

This work provides an overview of my dissertation in the research area of clustering in subspace projections. It is part of an emerging research community, as large and high dimensional databases where clusters are hidden in subsets of the given attributes are widely used in today's application scenarios. My thesis provides novel subspace clustering models which have shown to achieve enhanced clustering accuracy on benchmark databases. Furthermore, they all provide few but relevant subspace clusters such that users are able to review the output result sets. With our research on efficient processing schemes we have developed scalable algorithms applicable on large and high dimensional databases. For the overall research community our objective evaluation study on subspace clustering techniques provides the basis for future research. Please feel encouraged to use and extend our OpenSubspace framework for your own research.

Acknowledgments
This research was funded by the cluster of excellence on Ultra-high speed Mobile Information and Communication of the German Research Foundation grant EXC 89.

8. REFERENCES

[1] C. Aggarwal, J. Wolf, P. Yu, C. Procopiuc, and J. Park. Fast algorithms for projected clustering. In *SIGMOD*, pages 61–72, 1999.

[2] R. Agrawal, J. Gehrke, D. Gunopulos, and P. Raghavan. Automatic subspace clustering of high dimensional data for data mining applications. In *SIGMOD*, pages 94–105, 1998.

[3] I. Assent, R. Krieger, E. Müller, and T. Seidl. DUSC: Dimensionality unbiased subspace clustering. In *ICDM*, pages 409–414, 2007.

[4] I. Assent, R. Krieger, E. Müller, and T. Seidl. VISA: Visual subspace clustering analysis. *SIGKDD Explorations*, 9(2):5–12, 2007.

[5] I. Assent, R. Krieger, E. Müller, and T. Seidl. EDSC: Efficient density-based subspace clustering. In *CIKM*, pages 1093–1102, 2008.

[6] I. Assent, R. Krieger, E. Müller, and T. Seidl. INSCY: Indexing subspace clusters with in-process-removal of redundancy. In *ICDM*, pages 719–724, 2008.

[7] I. Assent, E. Müller, R. Krieger, T. Jansen, and T. Seidl. Pleiades: Subspace clustering and evaluation. In *ECML PKDD*, pages 666–671, 2008.

[8] K. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft. When is nearest neighbors meaningful. In *IDBT*, pages 217–235, 1999.

[9] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases. In *KDD*, pages 226–231, 1996.

[10] S. Günnemann, I. Färber, H. Kremer, and T. Seidl. CoDA: Interactive cluster based concept discovery. In *VLDB*, 2010.

[11] S. Günnemann, I. Färber, E. Müller, and T. Seidl. ASCLU: Alternative subspace clustering. In *MultiClust at KDD*, 2010.

[12] S. Günnemann, E. Müller, I. Färber, and T. Seidl. Detection of orthogonal concepts in subspaces of high dimensional data. In *CIKM*, pages 1317–1326, 2009.

[13] I. Joliffe. *Principal Component Analysis*. Springer, New York, 1986.

[14] K. Kailing, H.-P. Kriegel, and P. Kröger. Density-connected subspace clustering for high-dimensional data. In *SDM*, pages 246–257, 2004.

[15] H.-P. Kriegel, P. Kröger, M. Renz, and S. Wurst. A generic framework for efficient subspace clustering of high-dimensional data. In *ICDM*, pages 250–257, 2005.

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

[17] G. Moise, J. Sander, and M. Ester. P3C: A robust projected clustering algorithm. In *ICDM*, pages 414–425, 2006.

[18] E. Müller, I. Assent, S. Günnemann, T. Jansen, and T. Seidl. OpenSubspace: An open source framework for evaluation and exploration of subspace clustering algorithms in WEKA. In *OSDM at PAKDD*, 2009.

[19] E. Müller, I. Assent, S. Günnemann, R. Krieger, and T. Seidl. Relevant subspace clustering: Mining the most interesting non-redundant concepts in high dimensional data. In *ICDM*, pages 377–386, 2009.

[20] E. Müller, I. Assent, R. Krieger, S. Günnemann, and T. Seidl. DensEst: Density estimation for data mining in high dimensional spaces. In *SDM*, pages 173–184, 2009.

[21] E. Müller, I. Assent, R. Krieger, T. Jansen, and T. Seidl. Morpheus: Interactive exploration of subspace clustering. In *KDD*, pages 1089–1092, 2008.

[22] E. Müller, I. Assent, and T. Seidl. HSM: Heterogeneous subspace mining in high dimensional data. In *SSDBM*, pages 497–516, 2009.

[23] E. Müller, S. Günnemann, I. Assent, and T. Seidl. Evaluating clustering in subspace projections of high dimensional data. *PVLDB*, 2(1):1270–1281, 2009.

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

[25] C. Procopiuc, M. Jones, P. Agarwal, and T. Murali. A monte carlo algorithm for fast projective clustering. In *SIGMOD*, pages 418–427, 2002.

[26] K. Sequeira and M. Zaki. SCHISM: A new approach for interesting subspace mining. In *ICDM*, pages 186–193, 2004.

[27] B. Silverman. *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, London, 1986.

[28] I. Witten and E. Frank. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, USA, 2005.

[29] M. L. Yiu and N. Mamoulis. Frequent-pattern based iterative projected clustering. In *ICDM*, pages 689–692, 2003.