A Framework for Evaluation and Exploration of Clustering Algorithms in Subspaces of High Dimensional Databases

Emmanuel Müller [◆] Ira Assent [■] Stephan Günnemann [●]
Patrick Gerwert [●] Matthias Hannen [●] Timm Jansen [●] Thomas Seidl [●]

[◆]Karlsruhe Institute of Technology (KIT), Germany
emmanuel.mueller@kit.edu

[■]Aarhus University, Denmark
ira@cs.au.dk

[●]RWTH Aachen University, Germany
{guennemann, gerwert, hannen, jansen, seidl}@cs.rwth-aachen.de

Abstract: In high dimensional databases, traditional full space clustering methods are known to fail due to the curse of dimensionality. Thus, in recent years, subspace clustering and projected clustering approaches were proposed for clustering in high dimensional spaces. As the area is rather young, few comparative studies on the advantages and disadvantages of the different algorithms exist. Part of the underlying problem is the lack of available open source implementations that could be used by researchers to understand, compare, and extend subspace and projected clustering algorithms. In this work, we discuss the requirements for open source evaluation software and propose the OpenSubspace framework that meets these requirements. OpenSubspace integrates state-of-the-art performance measures and visualization techniques to foster clustering research in high dimensional databases.

1 Introduction

In recent years, the importance of comparison studies and repeatability of experimental results is increasingly recognized in the databases and knowledge discovery communities. VLDB initiated a special track on *Experiments and Analyses* aiming at comprehensive and reproducible evaluations (e.g. [HCLM09, MGAS09, KTR10, SDQR10]). The conferences SIGMOD followed by SIGKDD have established guidelines for repeatability of scientific experiments in their proceedings. Authors are encouraged to provide implementations and data sets. While these are important contributions towards a reliable empirical research foundation, there is still a lack of open source implementations for many state-of-the-art approaches. In this paper we present such an open source tool for clustering in subspaces of high dimensional data.

Clustering is an unsupervised learning approach that groups data based on mutual similarity [HK01]. In high dimensional spaces, subspace clustering and projected clustering identify clusters in projections of the full dimensional space.

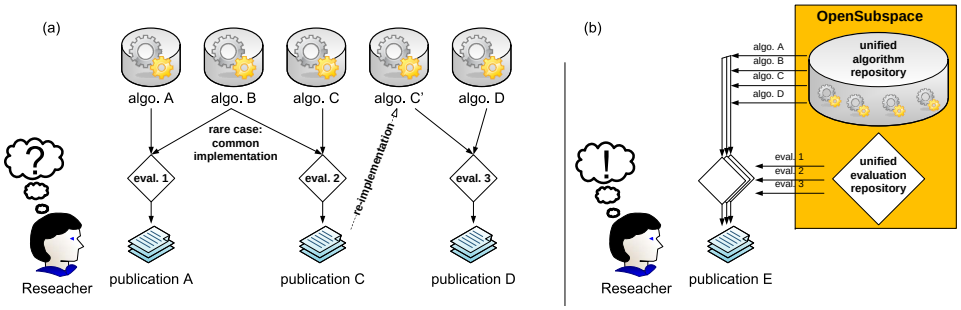

Figure 1: Subspace and projected clustering research with and without a general repository for data, algorithms, comparison, and evaluation.

It is a fundamental problem of unsupervised learning approaches that there is no generally accepted "ground truth". As clustering searches for previously unknown cluster structures in the data, it is not known a priori which clusters should be identified. This means that experimental evaluation is faced with enormous challenges. While synthetically generated data is very helpful in providing an exact comparison measure, it might not reflect the characteristics of real world data. In recent publications, labeled data, usually used to evaluate the performance of classifiers, i.e. supervised learning algorithms, is used as a substitute [SZ04, KKRW05, AKMS07a]. While this provides the possibility of measuring the performance of clustering algorithms, the base assumption that clusters reflect the class structure is not necessarily valid.

Some approaches therefore resort to the help of domain experts in judging the quality of the result [KKK04, BKKK04, KKRW05]. Where domain experts are available, which is clearly not always the case, they provide very realistic insights into the usefulness of a clustering result. Still, this insight is necessarily subjective and not reproducible by other researchers. Moreover, there is not sufficient basis for comparison, as the clusters that have not been detected are unknown to the domain expert. This problem is even more aggravated in high dimensional subspace or projected clustering. As the number of results is typically huge, it is not easily possible to manually analyze the quality of different algorithms or even different runs of the same algorithm.

As there is no ground truth, nor accepted benchmark data or measures for evaluating subspace and projected clustering, the experimental evaluation can be hardly set into relation to other published results. Especially the results are incomparable, as there are no publicly available common implementations neither for subspace/projected clustering algorithms nor for evaluation measures (cf. Fig 1). As a consequence, progress in this research area is slow, and general understanding of the advantages and disadvantages of different algorithms is not established. The source code for experimental evaluation is most of the time implemented by the authors themselves and often not made available to the general public. This hinders further experimental study of recent advances in clustering. As tedious re-implementation is often avoided, only few comparisons between new proposals and existing techniques are published.

For clustering (but also for classification and association rule mining), the open source tool WEKA (Waikato Environment for Knowledge Analysis) has been very helpful in allowing

researchers to analyze the behavior of different algorithms in comparison [WF05]. It provides measures for comparison, visualization of the results, and lets researchers add their own algorithms and browse through the implementation of other techniques.

For subspace and projected clustering, such a general tool does not exist. In this paper, we discuss the requirements for a successful open source tool that supports evaluation and exploration of subspace and projected clustering algorithms and their cluster results. Our framework OpenSubspace fulfills these requirements by integration of measurement and visualization techniques for in-depth analysis. Furthermore, it will be useful in establishing benchmark results that foster research in the area through better understanding of advantages and disadvantages of different algorithms on different types of data. It includes successful techniques in demonstration systems for visualization and evaluation of subspace mining paradigms [MAK+08, AMK+08, GFKS10, GKFS10, MSG+10].

As anyone will be able to see the implementation, the code base can be continually revised and improved. Researchers may analyze the algorithms on a far greater range of parameter values than would be possible within the scope of a single conference or journal publication (cf. Fig 1). Based on this, we published a thorough evaluation study on subspace/projected clustering techniques [MGAS09]. As open source basis for this study, this publication provides an overview of techniques included in our OpenSubspace framework.

For scientific publications the open source implementations in OpenSubspace enable more fine grained discussions about competing algorithms on a common basis. For authors of novel methods OpenSubspace gives the opportunity to provide their source code and thus deeper insight into their work. This enhances the overall quality of publications as comparison is not based any more on incomparable evaluations of results provided in different publications but on a common algorithm repository with approved algorithm implementations. In Figure 1 we compare the current situation (on the left side) with the improved situation having a common repository of both subspace/projected clustering and evaluation measures (on the right side). Thus, OpenSubspace aims at defining a common basis for research and education purposes maintained and extended by the subspace/projected clustering community.

None of the existing data mining frameworks provide both subspace/projected clustering as well as the full analytical and comparative measures for the full knowledge discovery cycle. KNIME (Konstanz Information Miner) is a data mining tool that supports data flow construction for knowledge discovery [BCD+09]. It allows visual analysis and integration of WEKA. Orange is a scripting or GUI object based component system for data mining [DZLC04]. It provides data modeling and (statistical) analysis tools for different data mining techniques. Rattle (the R Analytical Tool To Learn Easily) is a data mining toolkit that supports statistical data mining based on the open source statistical language R [Wil08]. Evaluation via a number of measures is supported. In all of these frameworks subspace clustering or projected clustering are not included. ELKI (Environment for DeveLoping KDD-Applications Supported by Index Structures) is a general framework for data mining [AKZ08]. While it also includes subspace and projected clustering implementations, the focus is on index support and data management tasks. With respect to evaluation and exploration, it lacks evaluation measures and visualization techniques for an easy comparison of clustering results. Furthermore, as a stand alone toolkit it does not provide an integration into popular tools like WEKA.

2 Subspace and Projected Clustering

Clustering is an unsupervised data mining task for grouping of objects based on mutual similarity [HK01]. In high dimensional data, the "curse of dimensionality" hinders meaningful clustering [BGRS99]. Irrelevant attributes obscure the patterns in the data. Global dimensionality techniques such as Principle Components Analysis (PCA), reduce the number of attributes [Jol86]. However, the reduction may obtain only a single clustering in the reduced space. For locally varying attribute relevance, this means that some clusters will be missed that do not show in the reduced space. Moreover, dimensionality reduction techniques are unable to identify clusterings in different reduced spaces. Objects may be part of distinct clusters in different subspaces.

Recent years have seen increasing research in clustering in high dimensional spaces. Projected clustering aims at identifying the locally relevant reduction of attributes for each object. More specifically, each object is assigned to exactly one cluster (or noise) and a corresponding projection. Subspace clustering allows identifying several possible subspaces for any object. Thus, an object may be part of more than one cluster in different subspaces.

2.1 Paradigms

While subspace and projected clustering are rather young areas that have been researched for only one decade, several distinct paradigms can be observed in the literature. Our open source framework includes representatives of these paradigms to provide an overview over the techniques available. We provide implementations of the most recent approaches from different paradigms (cf. Fig. 2):

Subspace clustering
Subspace clustering was introduced in the *CLIQUE* approach which exploits monotonicity on the density of grid cells for pruning [AGGR98]. *SCHISM* [SZ04] extends *CLIQUE* using a variable threshold adapted to the dimensionality of the subspace as well as efficient heuristics for pruning. Both are grid-based approaches which discretize the data space for efficient detection of dense grid cells in a bottom-up fashion.

In contrast, density-based subspace clustering defines clusters as dense areas separated by sparsely populated areas. In *SUBCLU*, a density monotonicity property is used to prune subspaces in a bottom-up fashion [KKK04]. *PreDeCon* extends this paradigm by introducing the concept of subspace preference weights to determine axis parallel projections [BKKK04]. A further extension *FIRES* proposes an approximative solution for efficient density-based subspace clustering [KKRW05]. In *DUSC*, dimensionality bias is removed by normalizing the density with respect to the dimensionality of the subspace [AKMS07a]. Its extension *INSCY* focuses on efficient in-process removal of redundancy [AKMS08]. Recently, more general techniques have been proposed for optimization of the resulting set of clusters to eliminate redundant results and to include novel knowledge in orthogonal projections [MAG+09, GMFS09].

Projected clustering
Projected clustering approaches are partitioning methods that identify disjoint clusters in

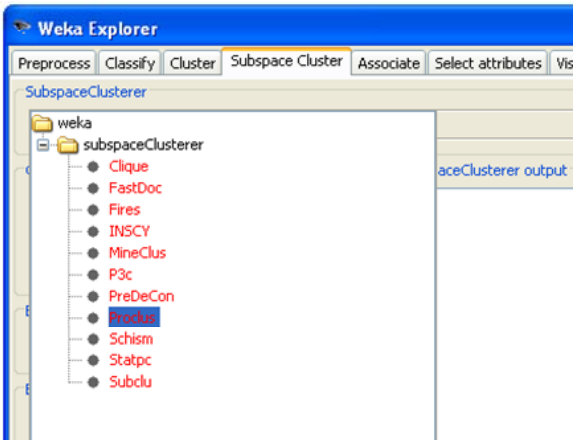

Figure 2: Algorithms implemented in OpenSubspace

subspace projections. *PROCLUS* extends the k-medoid algorithm by iteratively refining a full-space k-medoid clustering in a top-down manner [AWY+99]. *P3C* combines one-dimensional cluster cores to higher-dimensional clusters bottom-up [MSE06]. Its extension *StatPC* searches for non-redundant significant regions [MS08]. Further techniques are *DOC* a randomized approach using a Monte Carlo algorithm to find projected clusters represented by dense hypercubes [PJAM02] and *MineClus* an extension using the FP-tree for iterative projected clustering [YM03].

2.2 Challenges

Both subspace clustering and projected clustering pose new challenges to the mining task but especially to evaluation and exploration of the actual clustering results. In the following, we will show that these challenges have not yet been addressed by recent open source systems. Furthermore, they can not be solved by simply applying traditional techniques available for low dimensional clustering paradigms.

The WEKA framework provides several panels for different steps in the knowledge discovery cycle as well as for different data mining tasks (cf. Fig. 3). Besides structuring the GUI for users of the framework, the API reflects these different tasks in being structured according to classifiers, clustering algorithms, etc. This means that the Java class hierarchy reflects the common properties of each of the tasks. From this, several challenges arise in introducing a new data mining task, namely subspace/projected clustering, and new evaluation and visualization methods.

Due to special requirements in high dimensional mining we cannot simply extend the clustering panel in WEKA by adding new algorithms. We have to set up a new subspace panel by introducing techniques specialized to the new requirements in all areas (mining, evaluation and visualization). Subspace and projected clustering algorithms differ from clustering (or other data mining tasks such as classification) in that each cluster is associated with

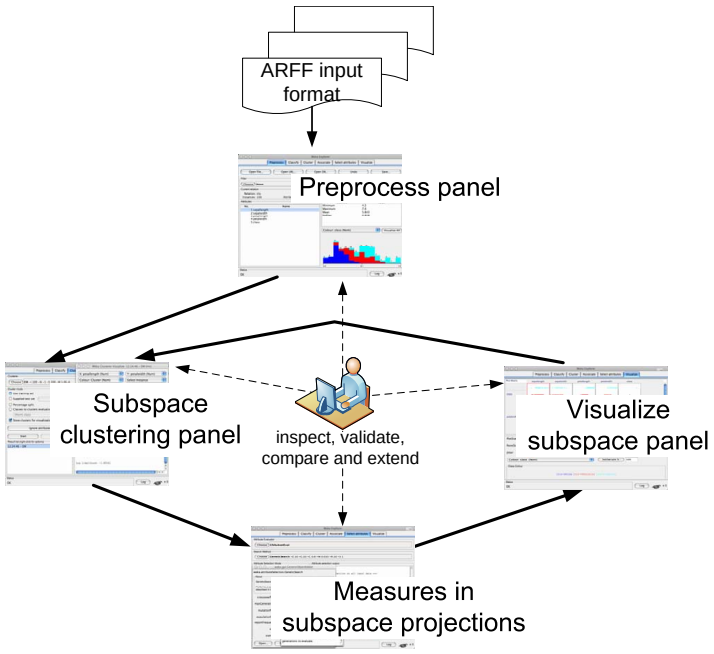

Figure 3: KDD cycle in WEKA for subspace and projected clustering

a possibly different subspace projection. As a consequence, the existing representation that assumes that all objects are clustered with respect to the initially chosen dimensions, is not valid. Moreover, the result is not necessarily partitioning. A single object may be part of several subspace clusters. These two aspects are important for the subspace/projected clustering panel, i.e. for the interface that describes common properties of these approaches. Moreover, these aspects have to be taken into consideration for evaluation and visualization as well. Following the same rationale, it is necessary to provide new APIs to allow meaningful analytical and visual tools in the OpenSubspace framework. Any measure that supports subspace and projected clustering evaluation needs to incorporate information on the respective subspace projection of each cluster. Visualization techniques to provide techniques for comparison of results in differing projections, as in [AKMS07b], cannot plug into existing visualization interfaces for traditional clustering in WEKA for the same reason.

As a consequence, OpenSubspace allows identifying any result with a corresponding set of dimensions, i.e. the subspace in which the result cluster resides. This is taken into consideration both for the subspace/projected clustering panel itself with the display of numeric results and evaluation measures, as well as for the visualization panel. This streamlined approach ensures that for all steps in the KDD cycle, representation in the correct subspace projection is achieved.

2.3 Lack of Ground Truth in Clustering

As briefly mentioned in the Introduction, clustering is a challenging task with respect to evaluation, as there is usually no ground truth. This means that it is in the very nature of clustering to search for yet unknown patterns in the data that provide novel and interesting insights to the user. Moreover, even for historic data, the true patterns that were interesting, are not known, either. This is in contrast to e.g. classification, where existing real world data can be used to easily validate the performance of any existing or newly presented classifier. Simply by checking the predicted class labels against the ones obtained from historic data, the classification accuracy can be easily measured in a reproducible fashion. In clustering, no such "labels" on historic data exists. Such "labels" would require an exhaustive enumeration of all combinatorial possibilities and their comparison. This is clearly infeasible even for reasonably small to medium datasets.

To allow for reproducible analysis, some publications resort to the measure of classification accuracy [BZ07, MAK+09]. The underlying idea is to find an objective measure for the performance of clustering. The assumption is that the class labels somehow reflect the natural grouping of the data, and can therefore be used to judge the performance of clustering algorithms as well. While this does provide some measure for comparison of these approaches, the underlying assumption is not necessarily valid and can even, in the worst case, produce random results. For example, an unsupervised clustering technique might detect a group of objects covering different labels, which might be meaningful as a clustering result. The given class labels, however, reflect only a single concept while clustering and especially subspace clustering aim at detecting multiple unknown concepts. The opposite case might happen as well: The clustering can split a set of objects with common labels into two clusters. Both meaningful clustering results are punished by evaluation measures simply based on the labels. As a consequence, class labels might provide only very limited insight into the performance of clustering algorithms.

Another approach taken in clustering evaluation is the use of synthetic data. Such artificial databases overcome the above mentioned problems by the generating process, the best clustering is already known. There are several limitations to this approach, however. First, most synthetic datasets are generated just for a single publication to evaluate the benefits of the proposed method. As such, they serve a very important purpose: they provide the means to understand whether the proposed method indeed detects (subspace or projected) clusters of the nature defined by the authors. Moreover, as the ideal clustering is known, the performance of algorithms perform on this dataset can be checked without having to resort to class labels. Even though some publications use very elaborate models to generate datasets that follow distributions that are believed to be observed in practical applications, there is obviously no guarantee that synthetic data is like real world data. Synthetic data, by its very nature, represents what is thought to occur in the datasets we analyze, but since we do not know which clusters might actually go unnoticed in real world data, these properties cannot be known.

Some researchers suggest using the help of domain experts in getting an informed answer to the quality of clustering results. Domain experts are obviously very helpful in judging the practical usefulness of the results and in ranking several possibilities in relation. However, as is true for the above examples, alternatives that are not known, i.e. not presented to the domain expert, cannot be taken into consideration. As a result, a very good cluster-

ing solution might be available and would be much more important to the domain expert. However, as this clustering is not retrieved, the domain expert cannot give a corresponding judgment. Moreover, manual analysis performed by domain experts is very limited to small result sizes and few parameter variations. Manual inspection of varied settings on a variety of datasets, as would be required for in-depth analysis, is clearly not feasible for humans. And, as mentioned before, the number of result clusterings in high dimensional spaces tends to grow enormously with the number of attributes. As a consequence, domain experts cannot judge typical outcomes of most subspace and projected clustering results. Moreover, the results indicated by domain experts are subjective and cannot be reproduced by other researchers.

As a consequence, any dataset used for evaluation is necessarily only a glimpse at the performance of (subspace or projected) clustering algorithms. As we will see later on, the open source idea provides a means to combine several of these glimpses into a larger picture towards an integrated view of clustering performance. As both the source code for validation of the results and the datasets are collected, a more integral picture is provided which can be easily extended by applying these algorithms to the datasets.

2.4 Lack of Standard Evaluation Measures in Clustering

Another problem in the evaluation of subspace and projected clustering lies in the evaluation measures themselves. This problem is closely related to the one of suitable datasets in that different results cannot be easily compared. Measuring the quality of (subspace and projected) clustering results is not straightforward. Even if the ground truth for any dataset were available, there are different ways of assessing deviations to this ground truth and of computing an overall performance score. For evaluation of clustering algorithms, large scale analysis is typically based on pre-labelled data, e.g. from classification applications [MSE06, BZ07]. The underlying assumption is that the clustering structure typically reflects the class label assignment. At least for relative comparisons of clustering algorithms, this provides measures of the quality of the clustering result.

In the literature, several approaches have been proposed. Quality can be determined as entropy and coverage. Corresponding roughly to the measures of precision and recall, entropy accounts for purity of the clustering (e.g. in [SZ04]), while coverage measures the size of the clustering, i.e. the percentage of objects in any subspace cluster. OpenSubspace provides both coverage and entropy (for readability, inverse entropy as a percentage) [AKMS07a]. Inverse entropy measures the homogeneity in the clustering result with respect to a class label. The measurement assumes a better clustering structure if the detected clusters are formed by objects homogeneously labeled with the same class labels. Besides the above mentioned problem of possible discrepancies in class labels and clustering structure, homogeneity of class labels is only one aspect of a good clustering structure. The coverage of the data set has to be measured separately to ensure that most of the objects occur in at least one cluster. Furthermore, overall homogeneity itself can be biased by many small homogeneous clusters dominating bigger inhomogeneous clusters.

Another approach is direct application of the classification accuracy. Accuracy of classifiers (e.g. C4.5 decision tree) built on the detected patterns compared with the accuracy of the same classifier on the original data is another quality measure [BZ07]. It indicates to

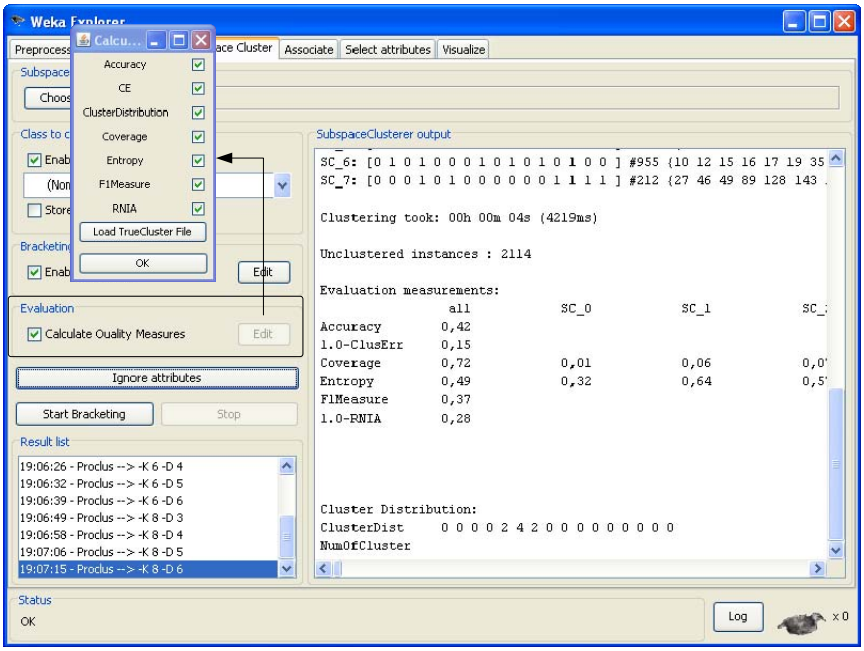

Figure 4: Evaluation measures in OpenSubspace

which extend the subspace clustering successfully generalizes the underlying data distribution. This approach is refined to enhanced classification accuracy which takes the original attributes and the ones that are derived as a combination of the original ones through the clustering. By comparing the performance of classifiers on the original attributes only with the performance of the same classifiers on original plus derived attributes, an insight into the quality of the clustering is achieved. However, as discussed before, any measured improvement is valid only with respect to the class labels. It is unclear, in which way the findings generalize to data without class labels.

The F1 value is commonly used in evaluation of classifiers and recently also for subspace or projected clustering as well [MSE06]. The F1 measure compares the clusters that are found by any particular (subspace or projected) clustering algorithm with an assumed ground truth by taking the harmonic mean of precision and recall. This approach obviously suffers from the same drawbacks as any class label-based method, yet, additionally, it is open to interpretation in high dimensional spaces. As clusters are detected in subspace projections, any deviation might be punished with respect to the subspace projection and with respect to the inclusion of positive and negative false alarms. However, the basis for comparison is not as straightforward, as it might seem. Depending on whether individual clusters or the entire clustering are used for the assessment, different results might be achieved. As a consequence, results based on variants of the F1 measure are not comparable across publications as one is using different F1 measure definitions.

All mentioned measures simply compare the detected groups of objects against the class label given for each object. Thus, these measures only provide a quality criterion for

object clusters as they ignore the detected subspaces in each cluster. More enhanced subspace clustering measures take also the detected subspaces into account and compare them against the possibly given relevant dimensions [PM06]. We have implemented such measures like Cluster Error (CE) and Relative Non-Intersecting Area (RNIA) in our framework. However, they require not only class labeled data for evaluation but also the relevant dimensions for each label. Such information is not provided in most real world data sets (e.g. used in classification task). Relevant dimensions are only available for synthetic data, as they are used and provided by the generators that hide subspace clusters in high dimensional spaces. Although both CE and RNIA achieve more detailed measurements as they take both objects and dimensions into account the missing ground truth is even more obvious for these measures.

In addition to this, several other measures have been used. In general, they all require some ground truth for assessing the performance of (subspace or projected) clustering algorithms. And, since several different subspace clusters might combine into a single "true" projected cluster, it is not always clear how to judge the result in its deviation from the postulated ground truth. Consequently, published results cannot be compared simply by their performance scores. Since there is no objective best measure for all approaches that is commonly agreed upon, researchers cannot compare different algorithms based solely on published results.

OpenSubspace provides the framework for using several, widely used, evaluation measures for subspace and projected clustering algorithms. In Figure 4 we present the evaluation output with various measures for comparing subspace clustering results. This allows easy extension of published results for various measures and direct comparison. Over time, as more and more results are available on different datasets and with respect to different evaluation measures, a benchmark background is built. It provides the means for in-depth understanding of algorithms and evaluation measures and fosters research in this area based on individual researcher's findings.

3 OpenSubspace Framework

With OpenSubspace we provide an open source framework tackling the challenges mentioned in the previous section. By fully integrating OpenSubspace into the WEKA framework we build on an established data mining framework covering the whole KDD cycle: pre-processing, mining, evaluation and visualization of the results, additionally including user feedback to the mining algorithm to close the KDD cycle. With OpenSubspace we focus on the mining, evaluation and exploration steps in this cycle (cf. Fig. 5). Providing a common basis for subspace/projected clustering as a novel mining step we achieve a framework for fair comparison of different approaches. For evaluation and exploration of the subspace and projected clusters OpenSubspace provides various evaluation measures for objective comparison of different clustering results. Furthermore, OpenSubspace provides visualization methods for an interactive exploration. Please refer to our website where we document our ongoing work in this project. It also contains more detailed information about OpenSubspace, its usage and extension. In the following, we will give an overview on the major contributions of OpenSubspace to the subspace clustering community:

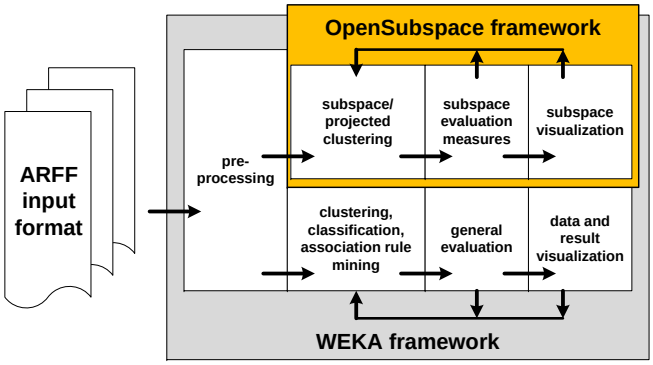

Figure 5: KDD cycle of OpenSubspace in WEKA

- Transparency of implementations
- Evaluation and comparison of algorithms
- Extensibility to further approaches

As we will show open source is the key property for all of these contributions. Open source code enables us to compare and validate the correctness of algorithm implementations. It gives us the basis for evaluating existing approaches on a common basis and leads thus to a fair comparison in future publications. By having the code of recent approaches at hand we enable the extension of existing algorithms to everyone and not only to the authors of these approaches.

3.1 Transparency of Implementations

The basis for thorough and fair evaluation is a common basis for all implementations. OpenSubspace provides such a basis for data access supporting both main and secondary storage. Furthermore, the framework provides a common interface for subspace clustering implementations. Algorithms which extend this interface can be easily plugged into OpenSubspace.

All of these algorithms are provided as open source. This transparency of the underlying implementation ensures high quality algorithms in the framework. The research community is able to review these implementations according to the original publication of the algorithm. Even improved versions can be provided which go beyond the descriptions in the publications using novel data structures, heuristics or approximation for specialized purposes. The benefit of reviewing implementations based on open source is especially useful as in most publications authors can only sketch their algorithms. This makes it difficult to re-implement such approaches. Various different interpretations of one approach could arise if only closed implementations were available. Using these different implementations of the same approach leads to incomparable results in scientific publications as evaluations have different bases. Open source repositories as in OpenSubspace prevent

such differing implementations that might even be biased and does away with the need for re-implementation for competing approaches. Overall OpenSubspace aims at a transparent and thus fair basis for evaluations of various approaches for detecting meaningful subspace clusters.

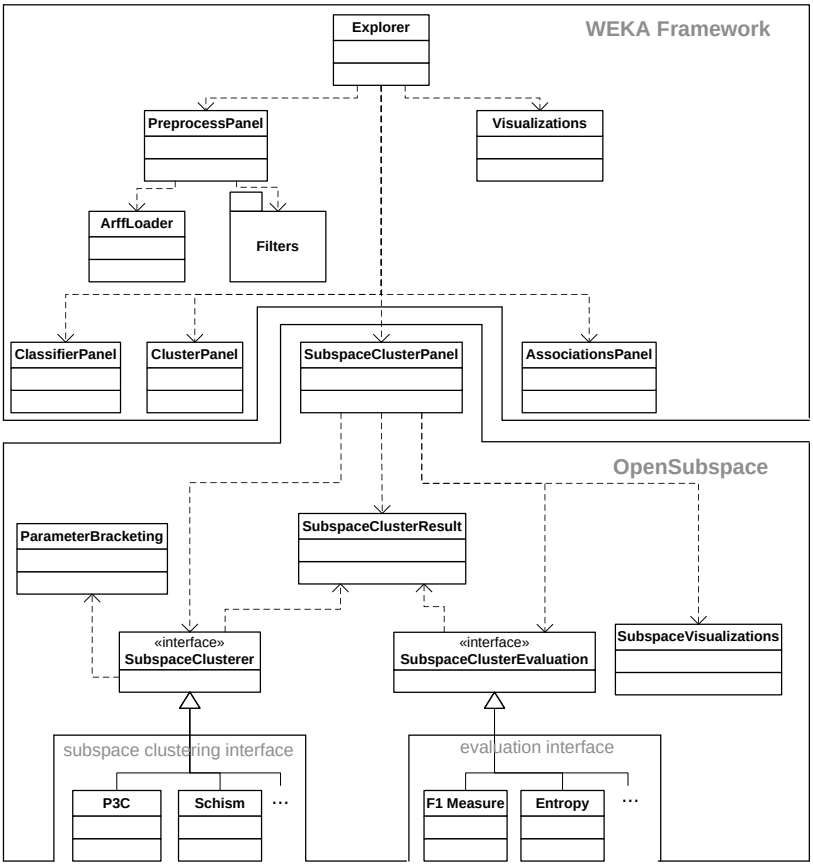

Figure 6: UML class diagram of the OpenSubspace framework

For extending the OpenSubspace algorithm repository our framework incorporates two open interfaces, which enable extensibility to further subspace clustering algorithms and new evaluation measurements. In Figure 6 we show the main classes of our OpenSubspace framework which extends the WEKA framework by a new subspace clustering panel.

Subspace clustering shows major differences compared to traditional clustering; e.g. an object can be part of several subspace clusters in different projections. We therefore do not extend the clustering panel, but provide a separate subspace clustering panel.

Recent subspace clustering algorithms described in Section 2.1 are implemented based on this framework. The abstraction of subspace clustering properties in OpenSubspace allows to easily add new algorithms through our new subspace clustering interface.

3.2 Evaluation and Comparison of Algorithms

Given the framework with transparent implementations of subspace clustering algorithms OpenSubspace enables researchers to evaluate their methods against competing approaches available in our repository. We establish a basis for developing new methods to perform an objective evaluation on arbitrary subspace clustering algorithms. OpenSubspace defines evaluation measurements based on labeled data sets. It includes measurements like entropy, coverage, F1-value, Cluster Error, RNIA and accuracy used in recent subspace/projected clustering publications as a basis for thorough evaluations. The set of measures is formally defened in our recent evaluation study providing additional experiments and analyses on several clustering paradigms [MGAS09].

In OpenSubspace all of these evaluation methods are implemented and published as open source as well. For a fair and comparative evaluation these measurements have to be accessible to all researchers. Review and refinement of these measurements is essential as there is always the possibility of different interpretations of these measures. As a ground truth is not given for subspace clustering the data mining community has to develop new evaluation measures that rate the quality of different approaches. This seems to be as difficult as the mining task itself. Therefore, we do not only provide several evaluation techniques (cf. Section 2.4) to measure the quality of the subspace clustering, but also an open interface (cf. Fig. 6) to implement new measures. Further measures can be added by our evaluation interface, which allows to define new quality criteria for subspace clustering on a common basis for all algorithms.

Evaluation measures summarize the result set in typically one real valued rating; however, visualization of results for more insight might be interesting. OpenSubspace, therefore, includes specialized visualizations for subspace clustering results with the possibility for interactive exploration. As stated before, subspace/projected clustering algorithms typically provide overwhelming result sets. Investigating these results is sometimes as difficult as looking at the raw data. For some specialized or domain dependent mining tasks it is even more important to investigate the actual clustering than to compare it with competing approaches. OpenSubspace provides specialized visualization techniques which close the KDD cycle by providing user feedback (cf. Fig. 5). Our framework provides interactive exploration of the results and thus the opportunity to refine the mining step by exploring different parameter settings and their resulting clustering output [MAK$^+$08, AMK$^+$08]. In addition the different views detected by subspace clustering approaches can be visualized and explored as well [GFKS10, GKFS10].

3.3 Visualization Techniques

OpenSubspace provides visualization techniques to present subspace clustering results such that users can easily gain an overview of the detected patterns, as well as an in-depth understanding of individual subspace clusters and their mutual relationship.

Gaining a meaningful overview is crucial in allowing users to assess the overall subspace clustering result. As mentioned, subspace clustering is inherently challenging as both the typical number of resulting subspace clusters is usually enormous as well as that clusters

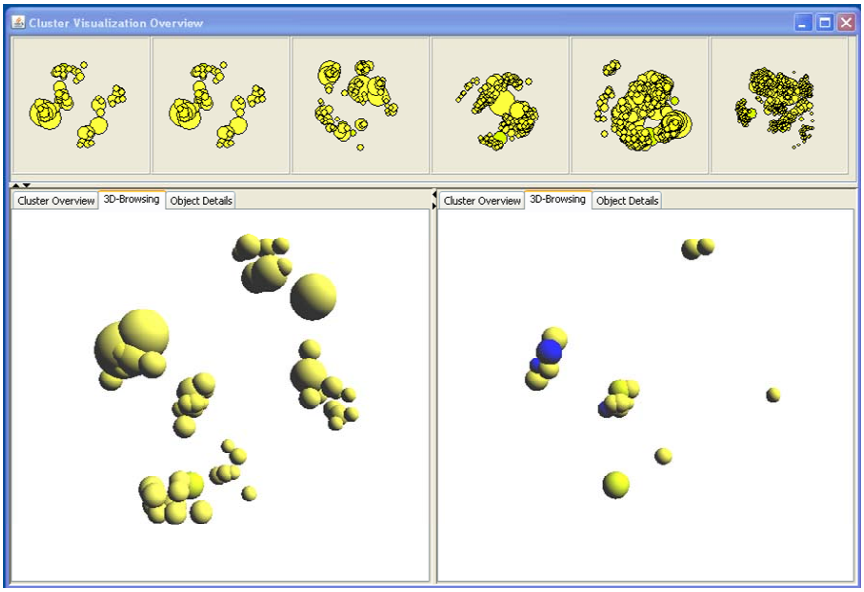

Figure 7: Visualization in OpenSubspace

in different projections are difficult to understand. Visualization techniques that were developed for full space clustering results rely on a common representation, i.e. no subspace projections [AKMS07b]. Consequently, they cannot be applied to subspace clustering.

Our framework thus contains specialized techniques for visualization of subspace clustering. 2d and 3d models are an adequate representation for human cognitive abilities. Based on a recently developed comparison measure for subspace clusters our system provides an overview on the entire subspace clustering result by MDS (multidimensional scaling) plots in both 2d and 3d [AKMS07b]. As illustrated in Figures 7, MDS approximates distances in high dimensional spaces by two or three dimensions. While the 2d representation is a static view that allows for easy reading, the 3d MDS plots allow users to interactively explore the overall subspace clustering result. They may move around the 3d representation to focus on those subspace clusters they are most interested in. At any point, they may choose individual subspace clusters in the plot to obtain more detailed information.

Thus, our MDS plots provide an overview on subspace clustering. Moreover, it helps users in interactively determining the best parameter setting. For any subspace clustering algorithm, some core parameters tend to have a large influence on the resulting output. We therefore present a bracketing representation, i.e. a series of 2d MDS plots for different parameters. Users thus get a clear visual impression of the effect of parameters and may choose the best ones for a feedback loop that generates the desired subspace clustering.

For in-depth analysis of any subspace clustering algorithm, representation of the key features of subspace clusters in a cognitively meaningful way is crucial. As subspace clustering results represent patterns in different projections by their very nature, visualization should contain information on the respective subspaces, the cluster values and additional

information on the interestingness measures computed by the subspace clustering algorithm. We use a color-coding scheme where the different axis in the HSV color space are used to represent different aspects of subspace clusters in a very compact and easy to understand manner [AKMS07b]. For easy navigation, subspace clusters can be zoomed into, and understood using a color legend on values of the subspace clusters.

3.4 Interactive Exploration

The conceptual design for interactive exploration in OpenSubspace is based on the *Visual Exploration Paradigm* [Kei02]: Starting from an overview over the subspace clustering result the user can navigate through the visualized patterns. By selecting subjectively interesting subspace clusters, the user may then obtain more detailed information where desired. Detailed information is provided on three levels: for entire subspace clusterings, for single subspace clusters, as well as for individual objects. Based on the discovered knowledge, the user can give feedback to the system for further improved results. This feedback loop enables the system to use the cognitive abilities of humans for better parameter settings and thus for meaningful subspace clusters.

Overview Browsing. Interactive exploration starts from an overview of all mined subspace clusters in which the user can browse. The automatically detected patterns are thus presented to the user for a general impression of the result and a comparison of the resulting clusters. As clusters are detected in arbitrary subspaces, they cannot be compared based on the full space dimensionality. We thus incorporate a distance function that takes the main characteristics of subspace clusters, their subspace dissimilarity and object dissimilarity into account for visualization in an MDS plot [AKMS07b]. Based on such an overall distance function, subspace clusters can be intuitively represented as circles in a 2d or 3d space (Figure 7). This approximation of the original high dimensional information to a 2d or 3d representation, allows human users to easily understand the result just by the visual impression. We enriches this MDS information by additional visual features. The size of a subspace cluster is represented as the diameter of the circle. Its dimensionality is encoded by the color of the circle. This information allows users to identify similar subspace clusters, those clusters of similar dimensionality, or of similar size, or to study the overall distribution of these characteristics in the result for further analysis.

Parameter Bracketing. Parameter setting is in general a difficult task for most data mining algorithms. This is especially the case for unsupervised techniques like clustering, where typically no prior knowledge about the data is available. This inherent problem of clustering is even more present in subspace clustering as the user has to provide parametrization for detecting clusters in different subspaces. In general the problem can be solved by guessing a parameter setting, looking at the result and then trying to choose a better parameter setting. To speed up this tedious process for users and give them more information to base their parameter choice on, we compute and visualize a series of different subspace clustering results at once, called bracketing of different parameter settings for direct feedback. This means that users obtain a series of MDS plots (cf. upper part of Figure 7) from which they pick the most appropriate one(s) for subsequent runs of the subspace clustering algorithms. By directly comparing the results of different parameter settings, parametrization is no longer a guess, but becomes an informed decision based

on the visual analysis of the effects of parameters. Moreover, this process is far more comfortable for users and allows reaching the desired subspace clustering result in fewer steps.

Direct subspace cluster comparison. For a more detailed analysis of two different parameter settings the user can select two clusterings out of the presented series of MDS plots by clicking on them in the bracketing representation. These two subspace clusterings are then presented as larger plots in the lower part of the cluster overview screen. Once again, detailed information for the subspace clusters can be obtain by picking individual subspace clusters.

3d Browsing. For the overview browsing we provide static 2-dimensional MDS plots. These static views provide a fixed perspective for easy comparison. For in-depth browsing, where focusing to different parts of the subspace clustering is of interest, a flexible navigation through MDS plots is provided. 3-dimensional MDS plot browsing allows users to shift, rotate and zoom into the MDS plot using standard 2-dimensional input devices or 3-dimensional input devices that allow for even more intuitive navigation. The user may thus identify similar or dissimilar subspace clusters that are of specific interest. In Figure 7, we show two 3-dimensional MDS plots representing two clustering results.

Interactive Concept Detection. In general, subspace clustering techniques were developed for the task of finding clusters in differing subspaces. Even more challenging is the grouping of clusters according to their specific concepts, for example the clusters 'smokers', 'joggers', or 'vegans' are manifestations of the concept 'health awareness'. Some recent approaches focus already on the task of grouping objects according to underlying concept structures [CFD07, GMFS09, GFMS10]: they find clusters in strongly differing subspace projections, providing the key for discovering the inherent concept structure. However, since the concepts are generative, i.e. they actually induce the clusters, they cannot be automatically concluded out of clusters. Accordingly, the mentioned subspace clustering techniques achieve concept-based aggregations of objects but are not capable of abstracting from these aggregations in the sense of named concepts.

In real-world applications, however, the interest lies in the explicit discovery and naming of the underlying concepts. This task cannot be solved automatically by unsupervised learning methods as subspace clustering but requires the domain knowledge of an expert. OpenSubspace supports the user in revealing the concepts out of a given subspace clustering [GFKS10, GKFS10]. It therefore provides the user with concept-oriented cluster visualization and interactive exploration to enable him to uncover the inherent concept structures. Each concept can be described by its occurring clusters on the one hand and its characteristic attributes on the other hand. Since the related clusters are not known beforehand, the idea is to capture the concepts through the structure of relevant attributes of the clustering. The relevant attributes are of particular importance for a semantic labeling of clusters and concepts. In the OpenSubspace framework, the user can take a closer look at the concept compositions and one can give feedback to refine or to recalculate the concept structures. Thus, the whole process of concept discovery in OpenSubspace is iterative and highly dependent on user interaction.

3.5 Extensibility of OpenSubspace

As a novel framework OpenSubspace provides the basis for further research. There are several algorithms implemented in our subspace/projected clustering repository. For evaluation measures we have included recently used measures in this field [SZ04, MSE06, BZ07, MAK+09, AKMS08]. However, as subspace clustering has just started to become a broader research topic, these evaluation measures can be only seen as first steps that are likely to be extended greatly in the near future. We included different visualization techniques in OpenSubspace which we presented in recent demonstration systems [MAK+08, AMK+08, GFKS10, GKFS10, MSG+10].

All three areas (mining, evaluation and visualization with interactive exploration) can be extended by open interfaces. Due to the fact that the whole framework is given as open source code it is easy to develop new algorithms, evaluation measures and visualizations. For researchers who wish to develop their own novel algorithm in this field we provide an easy way to integrate their approach into our framework and to perform a fair evaluation with competing approaches. Thus it is a key property of OpenSubspace to define an open basis for the development of new approaches, evaluation and visualization techniques.

We used and still use our framework for subspace clustering research but also for education in advanced data mining courses. In both cases we got positive feedback from our students who enjoyed easy and wide access and the predefined interfaces in our framework. Furthermore, we got encouraging feedback also by the community for our recent demonstration system which integrates extensible mining techniques into WEKA.

4 Conclusion

With OpenSubspace we provide an open source framework for the very active research area of subspace clustering and projected clustering. The aim of our framework is to establish a basis for comparable 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.

5 Ongoing and Future Work

OpenSubspace can be seen as the natural basis for our next task. We plan to develop 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. However, further research of meaningful and intuitive visualization is clearly necessary for subspace clustering. The open source framework for subspace mining algorithms has already encouraged researches in Visual Analytics and Human Computer Interaction to work on 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 high dimensional data sets. On such a benchmarking set one could collect best parameter settings for various algorithms, best quality results and screenshots of subspace clustering result visualizations 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/projected clustering research field in which one can easily judge the quality of novel algorithms by comparing it with approved results of competing approaches.

Acknowledgment

We would like to thank the authors of SUBCLU, FIRES and MineClus for providing us with the original implementations of their approaches, which we adapted to our framework.

References

[AGGR98] Rakesh Agrawal, Johannes Gehrke, Dimitrios Gunopulos, and Prabhakar Raghavan. Automatic subspace clustering of high dimensional data for data mining applications. *SIGMOD Record*, 27(2):94–105, 1998.

[AKMS07a] Ira Assent, Ralph Krieger, Emmanuel Müller, and Thomas Seidl. DUSC: Dimensionality Unbiased Subspace Clustering. In *Proc. IEEE International Conference on Data Mining (ICDM)*, pages 409–414, 2007.

[AKMS07b] Ira Assent, Ralph Krieger, Emmanuel Müller, and Thomas Seidl. VISA: visual subspace clustering analysis. *SIGKDD Explorations*, 9(2):5–12, 2007.

[AKMS08] Ira Assent, Ralph Krieger, Emmanuel Müller, and Thomas Seidl. INSCY: Indexing Subspace Clusters with In-Process-Removal of Redundancy. In *Proc. IEEE International Conference on Data Mining (ICDM)*, pages 719–724, 2008.

[AKZ08] Elke Achtert, Hans-Peter Kriegel, and Arthur Zimek. ELKI: A Software System for Evaluation of Subspace Clustering Algorithms. In *Proc. International Conference on Scientific and Statistical Database Management (SSDBM)*, pages 580–585, 2008.

[AMK+08] Ira Assent, Emmanuel Müller, Ralph Krieger, Timm Jansen, and Thomas Seidl. Pleiades: Subspace Clustering and Evaluation. In *Proc. European conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD)*, pages 666–671, 2008.

[AWY+99] Charu C. Aggarwal, Joel L. Wolf, Philip S. Yu, Cecilia Procopiuc, and Jong Soo Park. Fast algorithms for projected clustering. *SIGMOD Record*, 28(2):61–72, 1999.

[BCD+09] Michael R. Berthold, Nicolas Cebron, Fabian Dill, Thomas R. Gabriel, Tobias Kötter, Thorsten Meinl, Peter Ohl, Kilian Thiel, and Bernd Wiswedel. KNIME - the Konstanz information miner: version 2.0 and beyond. *SIGKDD Explorations*, 11(1):26–31, 2009.

[BGRS99] Kevin S. Beyer, Jonathan Goldstein, Raghu Ramakrishnan, and Uri Shaft. When Is "Nearest Neighbor" Meaningful? In *Proc. International Conference on Database Theory (ICDT)*, pages 217–235, 1999.

[BKKK04] Christian Böhm, Karin Kailing, Hans-Peter Kriegel, and Peer Kröger. Density Connected Clustering with Local Subspace Preferences. In *Proc. IEEE International Conference on Data Mining (ICDM)*, pages 27–34, 2004.

[BZ07] Björn Bringmann and Albrecht Zimmermann. The Chosen Few: On Identifying Valuable Patterns. In *Proc. IEEE International Conference on Data Mining (ICDM)*, pages 63–72, 2007.

[CFD07] Ying Cui, Xiaoli Z. Fern, and Jennifer G. Dy. Non-redundant Multi-view Clustering via Orthogonalization. In *Proc. IEEE International Conference on Data Mining (ICDM)*, pages 133–142, 2007.

[DZLC04] Janez Demar, Bla Zupan, Gregor Leban, and Tomaz Curk. Orange: From Experimental Machine Learning to Interactive Data Mining. In *Proc. Knowledge Discovery in Databases (PKDD)*, pages 537–539, 2004.

[GFKS10] Stephan Günnemann, Ines Färber, Hardy Kremer, and Thomas Seidl. CoDA: Interactive Cluster Based Concept Discovery. In *Proc. of the VLDB Endowment*, pages 1633–1636, 2010.

[GFMS10] Stephan Günnemann, Ines Färber, Emmanuel Müller, and Thomas Seidl. ASCLU: Alternative Subspace Clustering. In *Proc. MultiClust Workshop at ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2010.

[GKFS10] Stephan Günnemann, Hardy Kremer, Ines Färber, and Thomas Seidl. MCExplorer: Interactive Exploration of Multiple (Subspace) Clustering Solutions. In *Proc. IEEE International Conference on Data Mining (ICDM)*, 2010.

[GMFS09] Stephan Günnemann, Emmanuel Müller, Ines Färber, and Thomas Seidl. Detection of orthogonal concepts in subspaces of high dimensional data. In *Proc. ACM Conference on Information and Knowledge Management (CIKM)*, pages 1317–1326, 2009.

[HCLM09] O. Hassanzadeh, F. Chiang, H.C. Lee, and R.J. Miller. Framework for evaluating clustering algorithms in duplicate detection. *Proc. of the VLDB Endowment*, 2(1):1282–1293, 2009.

[HK01] Jiawei Han and Micheline Kamber. *Data Mining: Concepts and Techniques*. Morgan Kaufmann, 2001.

[Jol86] Ian Jolliffe. *Principal Component Analysis*. Springer, New York, 1986.

[Kei02] Daniel A. Keim. Information Visualization and Visual Data Mining. *IEEE Transactions on Visualization and Computer Graphics*, 8(1):1–8, 2002.

[KKK04] Peer Kröger, Hans-Peter Kriegel, and Karin Kailing. Density-Connected Subspace Clustering for High-Dimensional Data. In *Proc. SIAM International Conference on Data Mining (SDM)*, pages 246–257, 2004.

[KKRW05] Hans-Peter Kriegel, Peer Kröger, Matthias Renz, and Sebastian Wurst. A Generic Framework for Efficient Subspace Clustering of High-Dimensional Data. In *Proc. IEEE International Conference on Data Mining (ICDM)*, pages 250–257, 2005.

[KTR10] Hanna Köpcke, Andreas Thor, and Erhard Rahm. Evaluation of entity resolution approaches on real-world match problems. *Proc. of the VLDB Endowment*, 3(1):484–493, 2010.

[MAG+09] Emmanuel Müller, Ira Assent, Stephan Günnemann, Ralph Krieger, and Thomas Seidl. Relevant Subspace Clustering: Mining the Most Interesting Non-Redundant Concepts in High Dimensional Data. In *Proc. IEEE International Conference on Data Mining (ICDM)*, pages 377–386, 2009.

[MAK+08] Emmanuel Müller, Ira Assent, Ralph Krieger, Timm Jansen, and Thomas Seidl. Morpheus: interactive exploration of subspace clustering. In *Proc. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1089–1092, 2008.

[MAK+09] Emmanuel Müller, Ira Assent, Ralph Krieger, Stephan Günnemann, and Thomas Seidl. DensEst: Density Estimation for Data Mining in High Dimensional Spaces. In *Proc. SIAM International Conference on Data Mining (SDM)*, pages 173–184, 2009.

[MGAS09] Emmanuel Müller, Stephan Günnemann, Ira Assent, and Thomas Seidl. Evaluating clustering in subspace projections of high dimensional data. *Proc. of the VLDB Endowment*, 2(1):1270–1281, 2009.

[MS08] Gabriela Moise and Jörg Sander. Finding non-redundant, statistically significant regions in high dimensional data: a novel approach to projected and subspace clustering. In *Proc. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 533–541, 2008.

[MSE06] Gabriela Moise, Jörg Sander, and Martin Ester. P3C: A Robust Projected Clustering Algorithm. In *Proc. IEEE International Conference on Data Mining (ICDM)*, pages 414–425, 2006.

[MSG+10] Emmanuel Müller, Matthias Schiffer, Patrick Gerwert, Matthias Hannen, Timm Jansen, and Thomas Seidl. SOREX: Subspace Outlier Ranking Exploration Toolkit. In *Proc. European conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD)*, pages 607–610, 2010.

[PJAM02] Cecilia M. Procopiuc, Michael Jones, Pankaj K. Agarwal, and T. M. Murali. A Monte Carlo algorithm for fast projective clustering. In *Proc. ACM SIGMOD International Conference on Management of Data*, pages 418–427, 2002.

[PM06] Anne Patrikainen and Marina Meila. Comparing Subspace Clusterings. *IEEE Transactions on Knowledge and Data Engineering*, 18(7):902–916, 2006.

[SDQR10] Jörg Schad, Jens Dittrich, and Jorge-Arnulfo Quiané-Ruiz. Runtime Measurements in the Cloud: Observing, Analyzing, and Reducing Variance. *Proc. of the VLDB Endowment*, 3(1):460–471, 2010.

[SZ04] Karlton Sequeira and Mohammed Javeed Zaki. SCHISM: A New Approach for Interesting Subspace Mining. In *Proc. IEEE International Conference on Data Mining (ICDM)*, pages 186–193, 2004.

[WF05] Ian H. Witten and Eibe Frank. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, USA, 2005.

[Wil08] Graham Williams. Rattle: A graphical user interface for data mining in R using GTK. R package version 2.3.115. http://rattle.togaware.com/, 2008.

[YM03] Man Lung Yiu and Nikos Mamoulis. Frequent-pattern based iterative projected clustering. In *Proc. IEEE International Conference on Data Mining (ICDM)*, pages 689–692, 2003.