Pleiades: Subspace Clustering and Evaluation

Ira Assent, Emmanuel Müller, Ralph Krieger, Timm Jansen, and Thomas Seidl

Data management and exploration group, RWTH Aachen University, Germany
{assent,mueller,krieger,jansen,seidl}@cs.rwth-aachen.de
http://dme.rwth-aachen.de

Abstract. Subspace clustering mines the clusters present in locally relevant subsets of the attributes. In the literature, several approaches have been suggested along with different measures for quality assessment.

Pleiades provides the means for easy comparison and evaluation of different subspace clustering approaches, along with several quality measures specific for subspace clustering as well as extensibility to further application areas and algorithms. It extends the popular WEKA mining tools, allowing for contrasting results with existing algorithms and data sets.

1 *Pleiades*

In high dimensional data, clustering is hindered through many irrelevant dimensions (cf. "curse of dimensionality" [1]). Subspace clustering identifies locally relevant subspace projections for individual clusters [2]. As these are recent proposals, subspace clustering algorithms and respective quality measures are not available in existing data mining tools. To allow researchers and students alike to explore the strengths and weaknesses of different approaches, our *Pleiades* system provides the means for their comparison and analysis as well as easy extensibility.

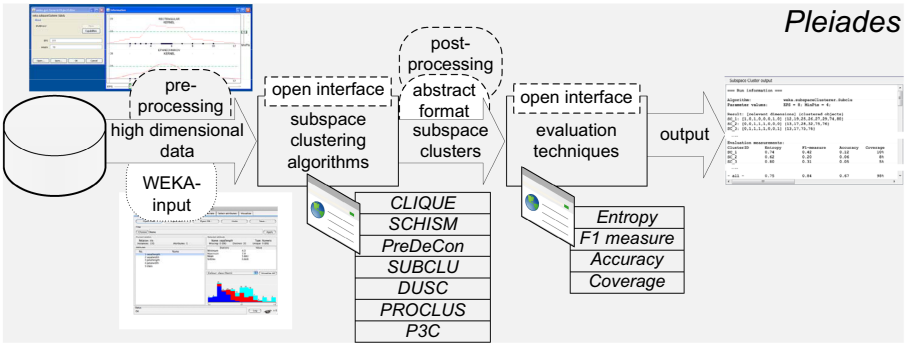

Fig. 1. Data processing in *Pleiades*

W. Daelemans et al. (Eds.): ECML PKDD 2008, Part II, LNAI 5212, pp. 666–671, 2008.

Figure 1 gives an overview over *Pleiades*: High-dimensional data is imported using the WEKA input format, possibly using pre-processing tools already available in WEKA and those we specifically added for subspace clustering [3]. The subspace clustering algorithms described in the next section are all part of the *Pleiades* system, and new algorithms can be added using our new subspace clustering interface. The result is presented to the user, and can be subjected to post-processing (e.g. filtering). *Pleiades* offers several evaluation techniques, as described in the next section, to measure the quality of the subspace clustering. Further measures can be plugged into our new evaluation interface.

2 *Pleiades* Features

Our *Pleiades* system is a tool that integrates subspace clustering algorithms along with measures designated for the assessment of subspace clustering quality.

2.1 Subspace Clustering Algorithms

While subspace clustering is a rather young area that has been researched for only one decade, several distinct paradigms can be observed in the literature. Our *Pleiades* system includes representatives of these paradigms to provide an overview over the techniques available. We provide implementations of the most recent approaches from different paradigms:

Grid-based subspace clustering discretizes the data space for efficient detection of dense grid cells in a bottom-up fashion. It was introduced in the *CLIQUE* approach which exploits monotonicity on the density of grid cells for pruning [4]. *SCHISM* [5] extends *CLIQUE* using a variable threshold adapted to the dimensionality of the subspace as well as efficient heuristics for pruning.

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 [6]. *PreDeCon* extends this paradigm by introducing the concept of subspace preference weights to determine axis parallel projections [7]. In *DUSC*, dimensionality bias is removed by normalizing the density with respect to the dimensionality of the subspace [8].

Projected clustering methods identify disjoint clusters in subspace projections. *PROCLUS* extends the k-medoid algorithm by iteratively refining a full-space k-medoid clustering in a top-down manner [9]. *P3C* combines one-dimensional cluster cores to higher-dimensional clusters bottom-up [10].

2.2 Evaluation Techniques

Quality of clustering or classification is usually measured in terms of accuracy, i.e. the ratio of correctly classified or clustered objects. For clustering, the "ground truth", i.e. the true clustering structure of the data, is usually not known. In fact, it is the very goal of clustering to detect this structure. As a consequence, clustering algorithms are often evaluated manually, ideally with the help of domain experts. However, domain experts are not always available, and they might

not agree on the quality of the result. Their assessment of the clustering result is necessarily only based on the result itself, it cannot be compared to the "optimum" which is not known. Moreover, manual evaluation does not scale to large datasets or clustering result outcomes.

For more realistic evaluation of clustering algorithms, large scale analysis is therefore typically based on pre-labelled data, e.g. from classification applications [10,11]. 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.

Our *Pleiades* system provides the measures proposed in recent subspace clustering publications. In Figure 2 we present the evaluation output with various measures for comparing subspace clustering results.

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 [5]), while coverage measures the size of the clustering, i.e. the percentage of objects in any subspace cluster. *Pleiades* provides both coverage and entropy (for readability, inverse entropy as a percentage) [8].

The **F1-value** is commonly used in evaluation of classifiers and recently also for subspace or projected clustering as well [10]. It is computed as the harmonic mean of recall ("are all clusters detected?") and precision ("are the clusters accurately detected?"). The F1-value of the whole clustering is simply the average of all F1-values.

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

Subspace Cluster output

```
=== Run information ===

Algorithm:              weka.subspaceClusterer.Subclu
Parameter values:       EPS = 8; MinPts = 4;

Result: [relevant dimensions] {clustered objects}
SC_1: [1,0,1,0,0,0,1,0] {12,19,25,26,27,29,74,80}
SC_2: [0,0,1,1,1,0,0,0] {13,17,28,32,75,76}
SC_3: [0,1,1,1,1,0,0,1] {13,17,75,76}
  ....
Evaluation measurements:
ClusterID       Entropy         F1-measure      Accuracy    Coverage
SC_1            0.74            0.42            0.12            10%
SC_2            0.62            0.20            0.06             8%
SC_3            0.80            0.31            0.05             5%
  ....
- all -         0.75            0.84            0.67            98%
```

Fig. 2. Evaluation screen

quality measure [11]. It indicates to which extend the subspace clustering successfully generalizes the underlying data distribution.

2.3 Applicability and Extensibility

Our *Pleiades* system provides the means for researches and students of data mining to use and compare different subspace clustering techniques all in one system. Different evaluation measures allow in-depth analysis of the algorithm properties. The interconnection to the WEKA data mining system allows further comparison with existing full space clustering techniques, as well as pre- and post-processing tools [3]. We provide additional tools for pre- and post-processing for subspace clustering. Figure 3 gives an example of our novel visual assistance for parameter setting in *Pleiades*.

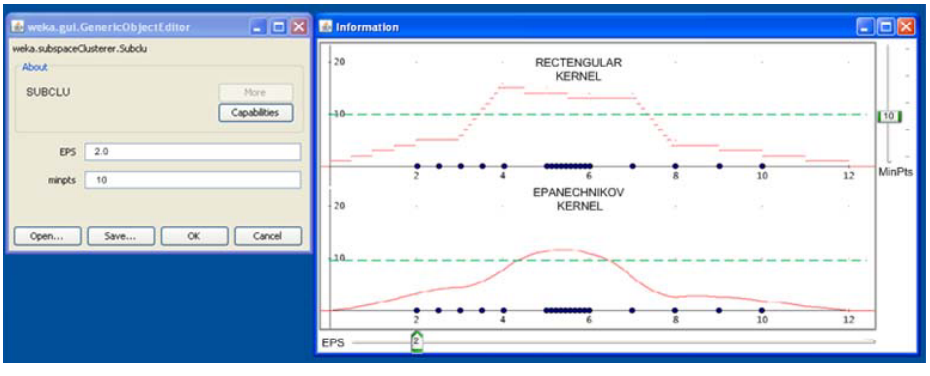

Fig. 3. Parametrization screen

Pleiades incorporates two open interfaces, which enable extensibility to further subspace clustering algorithms and new evaluation measurements. In Figure 4 we show the main classes of the *Pleiades* system 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 our *Pleiades* system. The abstraction of subspace clustering properties in *Pleiades* allows to easily add new algorithms through our new subspace clustering interface.

Furthermore, *Pleiades* offers several evaluation techniques (cf. Section 2.2) to measure the quality of the subspace clustering. By using these evaluation measures one can easily compare different subspace clustering techniques. Further measures can be added by our new evaluation interface, which allows to define new quality criteria for subspace clustering.

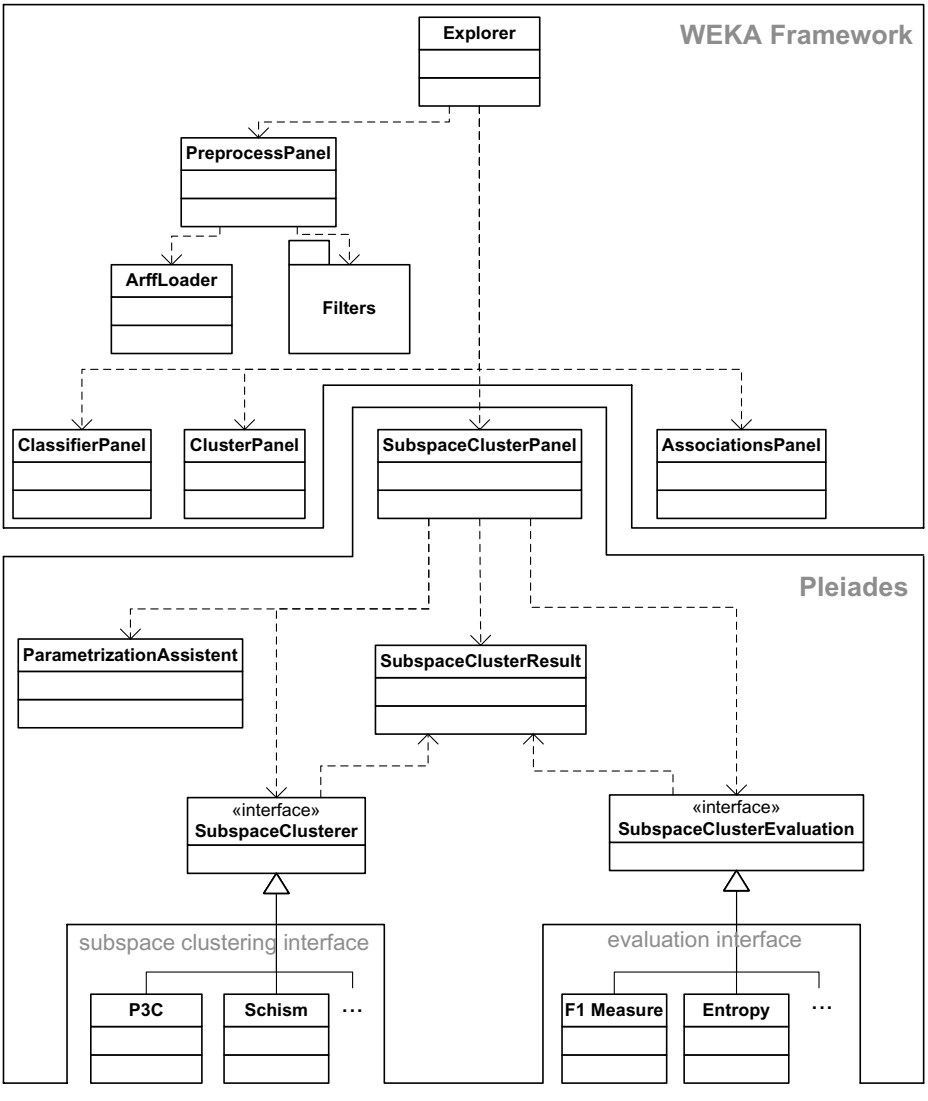

Fig. 4. UML class diagram of the *Pleiades* system

3 *Pleiades* Demonstrator

The demo will illustrate the above subspace clustering and evaluation techniques for several data sets. It will allow conference attendees to explore the diverse paradigms and measures implemented in the *Pleiades* system, thus raising research interest in the area. Open interfaces will facilitate extension with further subspace clustering algorithms and evaluation measures by other researchers.

References

1. Beyer, K., Goldstein, J., Ramakrishnan, R., Shaft, U.: When is nearest neighbors meaningful. In: Proceedings International Conference on Database Theory (ICDT), pp. 217–235 (1999)
2. Parsons, L., Haque, E., Liu, H.: Subspace clustering for high dimensional data: a review. ACM SIGKDD Explorations Newsletter 6(1), 90–105 (2004)
3. Witten, I., Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann Publishers Inc., San Francisco (2005)
4. Agrawal, R., Gehrke, J., Gunopulos, D., Raghavan, P.: Automatic subspace clustering of high dimensional data for data mining applications. In: Proceedings ACM SIGMOD International Conference on Management of Data, pp. 94–105 (1998)
5. Sequeira, K., Zaki, M.: SCHISM: A new approach for interesting subspace mining. In: Proceedings of the IEEE International Conference on Data Mining (ICDM), pp. 186–193 (2004)
6. Kailing, K., Kriegel, H.P., Kröger, P.: Density-connected subspace clustering for high-dimensional data. In: Proceedings of the IEEE International Conference on Data Mining (ICDM), pp. 246–257 (2004)
7. Böhm, C., Kailing, K., Kriegel, H.P., Kröger, P.: Density connected clustering with local subspace preferences. In: Proceedings of the IEEE International Conference on Data Mining (ICDM), pp. 27–34 (2004)
8. Assent, I., Krieger, R., Müller, E., Seidl, T.: DUSC: Dimensionality unbiased subspace clustering. In: Proceedings of the IEEE International Conference on Data Mining (ICDM), pp. 409–414 (2007)
9. Aggarwal, C., Wolf, J., Yu, P., Procopiuc, C., Park, J.: Fast algorithms for projected clustering. In: Proceedings ACM SIGMOD International Conference on Management of Data, pp. 61–72 (1999)
10. Moise, G., Sander, J., Ester, M.: P3C: A robust projected clustering algorithm. In: Proceedings of the IEEE International Conference on Data Mining (ICDM), pp. 414–425. IEEE Computer Society Press, Los Alamitos (2006)
11. Bringmann, B., Zimmermann, A.: The chosen few: On identifying valuable patterns. In: Proceedings of the IEEE International Conference on Data Mining (ICDM), pp. 63–72 (2007)