ASCLU: Alternative Subspace Clustering

Stephan Günnemann Ines Färber Emmanuel Müller Thomas Seidl

Data management and data exploration group
RWTH Aachen University, Germany
{guennemann, faerber, mueller, seidl}@cs.rwth-aachen.de

ABSTRACT

Finding groups of similar objects in databases is one of the most important data mining tasks. Recently, traditional clustering approaches have been extended to generate alternative clustering solutions. The basic observation is that for each database object multiple meaningful groupings might exist: the data allows to be clustered through different perspectives. It is thus reasonable to search for deviating clusters compared to a given clustering result, that the user is not satisfied with. The existing methods focus on full space clustering. However, for today's applications, where many attributes per object are recorded, traditional clustering is known to generate no meaningful results. Instead, the analysis of subspace projections of the data with subspace or projected clustering techniques is more suitable.

In this paper, we develop the first method that detects alternative *subspace* clusters based on an already known *subspace* clustering. Considering subspace projections, we can identify alternative clusters also based on deviating dimension sets besides just deviating object sets. Thus, we realize different views on the data by using different attributes. Besides the challenge of detecting alternative subspace clusters our model avoids redundant clusters in the overall result, i.e. the generated clusters are dissimilar among each other. In experiments we analyze the effectiveness of our model and show that meaningful alternative subspace clustering solutions are generated.

1. INTRODUCTION

Clustering is an established technique in the knowledge discovery process. It aims at detecting groups of similar objects while separating dissimilar ones. Traditionally, clustering algorithms 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 alternative groupings might exist for each object, e.g. in economics a customer can potentially belong to several customer groups. Various application scenarios like gene expression analysis, sensor networks, or the mentioned customer segmentation aim at detecting these multiple groups of objects.

The idea of multiple and possibly overlapping groupings of the objects requires different views on the data, which almost directly leads to the concept of subspace clustering. Here each object can participate in various alternative groupings, reflected in different subsets of the attributes as views on the same database. Different views on each object are thus achieved by considering (dis-)similarity based on subsets of attributes. Especially for data with very many attributes subspace clustering methods can still identify meaningful groupings based on locally relevant dimensions for each cluster.

Traditional clustering and subspace clustering methods do not act on the assumption that there exists some prior knowledge about groupings in the data. However, we might already know some trivial or already detected groupings in the data. If the user is not satisfied with the existent knowledge, either because it does not meet his application needs, or because he assumes there must exist more of those clusters in the data, then he aims for an alternative, yet comparable good clustering. In such scenarios the user is not willing to re-detect the already known clusters. As a general objective for recent alternative clustering techniques, it is important to acquire novel knowledge (not known in advance) by alternative clusters representing different views on the same database. The detection of such alternative clusters describing different views on each object is still an open challenge in recent applications.

Given a (subspace) clustering as prior knowledge, the task of alternative (subspace) clustering is to detect further alternative groupings hidden in different views of the given database. For example, in sensor analysis one aims at detecting sensor groups showing similar measurements. Each sensor might be grouped in multiple alternative clusters. One object might be clustered due to its high temperature and low humidity measurements in the "hot and dry region" cluster, while the same object might be clustered in the "light region" cluster considering only the illumination attribute. Assuming these two clusters as given prior knowledge, further interesting alternative clusters might be hidden in the database, e.g. a grouping of sensors representing a "dark and humid region". Such an alternative cluster might be of great importance in addition to the given two clusters. However, there might also be some trivial useless clusters, like objects clustered in both a "dry region" and in a "hot region". Obviously these two clusters only provide redun-

dant information to the given "hot and dry region" cluster. As illustrated in this toy example, detection of alternative clusters is of great importance, especially in recent applications where clusters are hidden in any possible attribute combination.

In general, we detect clusters hidden in subspace projections of the database to identify multiple views on the data. However, several new challenges arise for the research area of alternative subspace cluster detection. As one searches for clusters in arbitrary subspaces each cluster might be detected in multiple redundant views. Similarly, the knowledge of already given clusters might be repeated in similar subspace clusters. In both cases our novel approach aims at detecting new clusters not yet detected by other subspace clusters and not yet represented by the given clusters. Thus, our main contributions include:

- Detection of alternative subspace clusters
- Non-redundant clusters (dissimilar to each other)
- Alternative clusters (dissimilar to given clusters)

2. RELATED WORK

In this section, we review traditional and recent clustering techniques and show the differences to alternative clustering. Especially, we compare the detection of clusters in one fixed space and the consideration of multiple possible subspaces. This highlights the novelty of our approach, as some alternative clustering techniques have been introduced for one fixed space, while our approach focuses on alternative clustering in multiple subspace projections of the data.

Clustering without alternatives.

For the clustering techniques without alternative cluster detection one has to distinguish between traditional and subspace clustering. Traditional clustering approaches aim at the detection of clustered objects using all attributes in the full data space [11]. As in many applications clusters do not appear across all attributes, clustering has to focus on meaningful subsets of attributes. Since the relevance of attributes is usually not globally uniform for all clusters, global dimensionality reduction is inappropriate.

In research recently a number of different subspace clustering approaches [16, 12] has been introduced and are evaluated in [15]. *Subspace clustering* aims at detecting 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. Beside different cluster definitions, the major difference to traditional clustering can be seen in the detection of multiple clusters for each object. Subspace clustering allows objects to be part of multiple clusters in arbitrary subspaces [1, 2, 13].

As one of the most recent approaches in this area we have extended subspace clustering to the detection of orthogonal subspace clusters [9]. Aiming at the detection of knowledge in orthogonal subspaces, we reduce the exponential result set of traditional subspace clustering to only few but meaningful clusters. Thus, for each object one detects multiple alternative clusters in orthogonal subspaces. However, as for traditional clustering as well as for all other subspace clustering methods we do not assume to have a given clustering. Thus, results may lead to several already known clusters.

Alternative cluster detection.

Recent extensions of traditional clustering techniques try to detect alternative clusters to a given, known clustering. The techniques of [4, 5, 17, 3] base on a given clustering and iteratively transform the data space to force the underlying traditional clustering algorithm to find new, alternative clusters. Other techniques, like [8], follow the idea of using the conditional information bottleneck approach to find alternative clusterings. All these techniques are not able to detect clusters hidden in arbitrary subspace projections of the data and consider in each step only one fixed space. Furthermore, their input clustering has to be a partition of the data in a fixed space, whereas we allow a subspace clustering as input, which has multiple locally relevant subspaces. Only two of these approaches briefly refer to subspace clustering.

Although the method presented in [17] searches for clusters in the full space, it can be adapted to handle subspace clusters (of a single fixed subspace) as input by simply setting the values in the relevant attributes to zero. These dimensions therefore loose their influence in the following iterations. However, this complete elimination of covered attributes leads to an orthogonal subspace, which is a too strong restriction for the choice of relevant dimensions.

The approach in [4] has originally not been introduced to find alternatives for a given clustering but can easily be adapted by replacing the initial k-means clustering through the known clustering solution. Since this approach projects the data into an orthogonal space to find alternative clustering views, it is also not a subspace clustering and suffers from the problems already mentioned for [17].

Although these methods are, with large restrictions, able to find clusters in a (fixed) subspace, they are mostly not aware of the relevant subspaces and can therefore not annotate them to the clusters. The reason why objects group in a certain manner, however, originates from the respective subspace, which makes the relevant attributes an essential aspect to the clusters information.

3. ALTERNATIVE SUBSPACE CLUSTERS

In this section we describe our model for finding alternative subspace clusterings. To achieve this goal, our model adapts techniques of the OSCLU model [9] that finds orthogonal subspace clusterings in the data. With our novel method, however, we specifically address the problem of finding an alternative subspace clustering given a previously known subspace clustering. Thereby, we achieve that the user can steer the clustering algorithm to patterns not yet detected and the generation of already known clusters is prevented.

In general, a subspace cluster $C = (O, S)$ is a set of objects $O \subseteq DB$ and a set of dimensions $S \subseteq Dim$. The objects O are similar within the relevant dimensions S while the dimensions $Dim \backslash S$ are irrelevant for the cluster. In Fig. 1 the cluster C_1 corresponds to a 2-dimensional cluster while C_2 is a 1-dimensional one. The input of our model is an already known subspace clustering $Known = \{K_1, \ldots, K_m\}$ where K_i is a subspace cluster. The task is to identify another subspace clustering within the database that differs from the given one. For our example in Fig. 1 we assume $Known = \{C_1, C_2\}$. A possible alternative solution is $\{C_4, C_5, C_6\}$. This solution is interesting because we detect clusters in novel subspaces of the database. We designed our model to be independent of the actual cluster definition,

i.e. we assume a set $All = \{C_1, \ldots, C_k\}$ of possible subspace clusters is given (cf. Sec. 3.3 for our instantiation). In Fig. 1 we assume $All = \{C_1, \ldots, C_7\}$. The set All is also a subspace clustering; however, it is not an (good) alternative to $Known$. Besides others, this set contains clusters very similar to clusters in the input clustering. Thus, the overall goal of our model is to select a meaningful subset $Res \subseteq All$ as the result presented to the user.

Problem statement: Given an already known subspace clustering $Known = \{K_1, \ldots, K_m\}$, the aim of alternative subspace clustering *is to determine a meaningful subset $Res \subseteq All$ of all possible subspace clusters $All = \{C_1, \ldots, C_k\}$, such that Res differs from the input clustering.*

In the following we discuss the criteria a meaningful alternative clustering solution has to fulfill and we define the overall result.

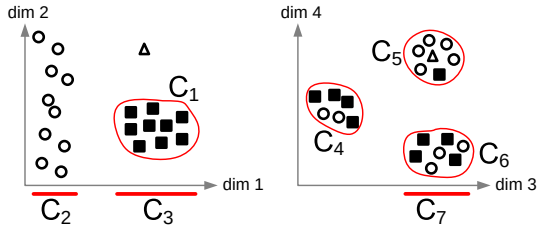

Figure 1: Exemplary subspace clusters

3.1 Valid alternative subspace clustering

Given the clustering $Known$, we want to detect a valid alternative clustering Res. Which properties must hold true for Res to be a *valid* alternative? Apparently, each cluster $C \in Res$ should considerably deviate from the clusters in $Known$. The cluster C should provide us with novel knowledge. For subspace clustering we have two possibilities to realize a deviation to already known clusters. First our novel subspace cluster comprises a "different" (i.e. novel) subspace or it covers "different" (i.e. novel) objects in already known subspaces. Thus, a cluster C is not a valid alternative if the subspace as well as the objects are already clustered.

Alternative w.r.t. subspaces. If the subspaces of two clusters differ substantially, both are interesting, even if their clustered object sets are nearly identical. Different subspaces mean different relevant attributes and hence a valid alternative. However, it is problematic to deduce that a cluster $C = (O, S)$ is a valid alternative to $C' = (O', S')$ only based on the fact $S \neq S'$. It is a well known observation in the area of subspace clustering that similar object groupings appear in very similar subspaces several times: this is one aspect of the redundancy problem in subspace clustering [14, 13, 2]. Considering for example the clusters C_1 and C_3 in Fig. 1, their subspaces are unequal but very similar. Thus, grouping of similar objects is expected. For our task of finding alternatives we have to ensure that the subspaces of our novel clusters differ to a high extend compared to the ones of the given input clusters. If a cluster $C \in Res$ and a cluster $K \in Known$ have highly deviating subspaces, we thus do not have to enforce deviating object sets for these two clusters: C is already a good alternative with respect to the single cluster K. The subset of clusters of $Known$ that are already different enough due to their subspaces is defined by:

DEFINITION 1. *Clusters in alternative subspaces.*
Given a cluster $C \in Res$, the subset of clusters of $Known$ that belong to an alternative subspace w.r.t. $C = (O, S)$ is defined by

$$InAlterSubspace(Known, C) = \{(O_i, S_i) \in Known \mid |S \cap S_i| < \beta \cdot |S|\}$$

with $0 < \beta \leq 1$.

If the fraction $|S \cap S_i|$ of the joint dimensions compared to all dimensions of C is small enough, the clusters represent different concepts and hence alternative information of the data. Thus, to decide whether C is a valid alternative to all clusters in $Known$, we can already neglect the clusters in $InAlterSubspace(Known, C)$. In Fig. 1 we get $InAlterSubspace(Known, C_4) = \{C_1, C_2\}$, because all of the input clusters were detected in completely alternative subspaces. For C_3 and with $\beta = 0.5$, however, we get $InAlterSubspace(Known, C_3) = \emptyset$.

Keep in mind that this relation is not symmetric. Assuming an input cluster K from $Known$ in the subspace $\{1, 2, 3\}$ and a novel identified cluster C in the subspace $\{2, 3, 4, 5, 6\}$, then $K \in InAlterSubspace(Known, C)$ for $\beta = 0.5$. The clusters share just two dimensions, which is significantly smaller compared to all five dimensions of C. Thus, C is already an alternative to K because there are enough new dimensions in C to provide novel information. However, assuming C is in $Known$ and K is the newly identified cluster, then $C \notin InAlterSubspace(Known, K)$. There are still two common dimensions but these are now compared to just the three dimensions of K. With respect to the subspace, K is not an alternative to C because K mainly has dimensions already being included in C. Thus, C is not in a different but in a similar subspace w.r.t. K.

Alternative w.r.t. objects. For the possible case of $InAlterSubspace(Known, C) \neq Known$ we have given some clusters that were detected in subspaces similar to C. Thus, for these clusters we have to ensure a grouping of different objects compared to C. The other clusters are already neglected because of deviating subspaces. As a criterion for deviating object representation we use the coverage of objects. If the covered objects of $C = (O, S)$ sufficiently differ from the covered objects of the clusters, the cluster C is a valid alternative. The already clustered objects of the clusters in similar subspaces are given by

$$AlreadyClustered(Known, C) = \bigcup_{(O,S)=K \in Known} \{O \mid K \notin InAlterSubspace(Known, C)\}$$

and we define:

DEFINITION 2. *Valid alternative subspace clustering*
Given a cluster $C \in Res$, $C = (O, S)$ is a valid alternative cluster to $Known$ iff

$$\frac{|O \setminus AlreadyClustered(Known, C)|}{|O|} \geq \alpha$$

with $0 < \alpha \leq 1$.
Given a clustering $Res \subseteq All$, Res is a valid alternative clustering to $Known$ iff all clusters $C \in Res$ are valid alternative clusters to $Known$.

In Fig. 1 the set $AlreadyClustered(Known, C_3)$ contains nearly all objects of C_3 itself. Thus, the fraction of objects not covered of C_3 is very low and, for example by choosing $\alpha = 0.4$ the cluster is not a valid alternative (less than 40% of the objects are novel). If the input clustering, however, is $Known = \{C_2, C_5\}$, the cluster C_3 is a valid alternative. C_5 is located in a different subspace, hence irrelevant for C_3, and C_2 covers different objects than C_3. Choosing $Known = \{C_2, C_5\}$, the clustering $Res = \{C_3, C_4, C_6\}$ is a valid alternative since either different subspaces or different objects are identified.

3.2 Optimal alternative subspace clustering

With Def. 2 we can find meaningful alternatives to the input clustering. However, several different alternatives exist as, e.g. $Res = \{C_3, C_4, C_6\}$ or $Res' = \{C_1, C_3, C_4, C_6\}$ in our previous example with $Known = \{C_2, C_5\}$. These solutions are not equally interesting for the user. In the following we demonstrate the difference in the solutions and we define the optimal alternative clustering.

Redundancy. A valid alternative clustering ensures that each cluster $C \in Res$ is different enough to the clusters in $Known$. Up to now, we do not set up further properties between the clusters of Res itself. Since for subspace clustering a partitioning of the objects is not enforced, the solution could contain very similar clusters. The solution $Res' = \{C_1, C_3, C_4, C_6\}$ is a valid alternative, although the clusters C_1 and C_3 are very similar to each other. Compared to the input clustering this is not a problem; however, for the solution itself we introduce redundancy. We should avoid the selection of both clusters simultaneously because only one of them provides novel knowledge. With our previously introduced definition we can solve this problem in an elegant and easy way. We require for each cluster $C \in Res$ to be different enough to all its remaining clusters in Res. That is, we have to ensure that C is a valid alternative to $Res \backslash \{C\}$. Thereby we cannot have redundant clusters in Res. In our example the set $\{C_3, C_4, C_6\}$ fulfills this property. Additionally including C_1, however, leads to redundancy: C_1 is not a valid alternative to the remaining clusters $\{C_3, C_4, C_6\}$.

Local interestingness. By avoiding redundant clusterings we reduce the number of possible clustering solutions. Nonetheless, many clusterings are still possible. As mentioned above, a simultaneous selection of C_1 and C_3 is not possible. However, selecting either C_1 or C_3 leads to two different results. To select the most interesting clustering result among all others, we use an idea also used in OSCLU. Each cluster is annotated with a certain interestingness value. By selecting those clusters that maximize this interestingness, we get the desired result. Formally, we have to define an interestingness function I that maps each subspace cluster C to the interestingness value $I(C)$. This function can incorporate several characteristics as the dimensionality or the size of the cluster. Our instantiation is presented in Sec. 3.3. After defining the interestingness function, the overall interestingness of a clustering Res is obtained by summing up the individual values of each cluster:

$$quality(Res) = \sum_{C \in Res} I(C)$$

Assuming higher dimensional clusters are more interesting in our example, the quality of $\{C_1, C_4, C_6\}$ is higher than the one of $\{C_3, C_4, C_6\}$.

Accounting for the redundancy and the local interestingness we are now able to define our overall clustering solution:

DEFINITION 3. *Optimal alternative subspace clustering. Given a previously known subspace clustering $Known$ and the set of all possible subspace clusters All, a clustering $Res \subseteq All$ is an optimal alternative subspace clustering iff*

a) Res is a valid alternative to $Known$

b) $\forall C \in Res : \{C\}$ is a valid alternative to $Res \backslash \{C\}$

c) Res is the most interesting clustering, i.e. $\forall Res' \subseteq All$ that also fulfill a & b: $quality(Res) \geq quality(Res')$

With this new model we are able to determine a subspace clustering result that differs from the input clustering to a high extent: either by representing novel objects or comprising novel subspaces. At the same time we avoid generating redundant clusters for the result, focusing again on deviating subspace clusters. Overall, we identify a meaningful alternative to the given subspace clustering.

3.3 Instantiation and algorithm

Instantiation. Our optimal alternative subspace clustering model is able to determine a meaningful clustering based on a given set of subspace clusters. To define this set of clusters we use the density-based clustering paradigm because it detects arbitrary shaped clusters even in noisy data [6]. A cluster is determined via dense areas in the data space [10]. As in OSCLU, the density $density^S(p)$ of a point p in subspace S is determined by the cardinality of its ε-neighborhood and the variable ε is adjusted to the dimensionality of the subspace [9]. Besides the cluster definition we also have to set up the interestingness function I. We use the identical approach as in OSCLU and incorporate the dimensionality, size and density of the corresponding subspace cluster $C = (O, S)$ in our function:

$$I(C) = |S|^a \cdot |O|^b \cdot \left(\frac{1}{|O|} \sum_{p \in O} density^S(p) \right)^c$$

with $a + b + c = 1$.

Overview of the algorithm. We will give a brief overview of the algorithm. The OSCLU model was proven to be NP-hard. Thus, for our model as well, we cannot expect that an efficient algorithm, exactly solving our model, exists. Instead, we develop an approximation algorithm that avoids generating the set of all possible subspace clusters by pruning several subspaces based on already detected patterns and using the knowledge of the input clustering. We incrementally add clusters to the current result set Res and we possibly refine this set if better clusters are detected. Technically, we use a top-down approach starting in high-dimensional spaces and traversing the subspace lattice in breadth-first order. During this traversal, the subspaces with the same dimensionality are processed in the order of their possible benefit for the clustering result. Subspaces that are highly different to subspaces already present in the input clustering $Known$ and different to the ones in the current result set are analyzed first. Too similar subspaces are pruned. The best ranked subspace is analyzed for its clustering structure using the density-based clustering model.

If clusters were identified in the current subspace, we check if these clusters can be added to the set Res. We have differ-

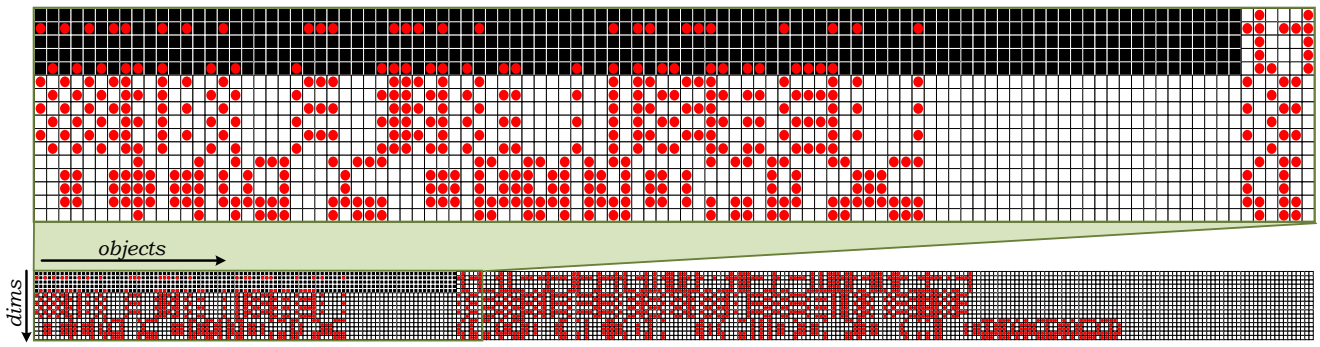

Figure 2: Data matrix for synthetic data with given clustering (black) and alternative clustering (red)

ent criteria: First, if the novel cluster C is not a valid alternative to $Known$, we can directly skip this cluster. Second, if C is a valid alternative to $Known$ and also to the current result set Res, we can directly add this cluster to the result, i.e. $Res = Res \cup \{C\}$. Last, if C is an alternative to $Known$ but not to Res, clusters $M \subseteq Res$ have to exist that are responsible for the (current) redundancy of C w.r.t. Res. If we removed M from Res, we could add C to the result. To maximize the interestingness of our clustering solution we determine the values $quality(Res \setminus M \cup \{C\})$ as well as $quality(Res)$ and we select the more interesting clustering.

After checking each cluster within the current subspace, the result set has possibly changed, thus, the order of the subspaces, not yet analyzed, is potentially adapted. We do not want to analyze subspaces that are similar to the ones in $Known$ and Res. Overall, we steer our algorithm only to those subspaces where alternative clusters are expected and we avoid analyzing all subspaces.

4. EXPERIMENTS

In this section, we evaluate the quality of the ASCLU (Alternative Subspace Clustering) approach and investigate whether it can provide a reasonable and non-redundant subspace clustering compared to a given set of input subspace clusters. For this goal we start by applying ASCLU to synthetic data to get a better intuition of the main principle. The generation method for synthetic data sets corresponds to the one used in [9]. It produces density-based clusters in arbitrary subspaces, where each object can belong to multiple clusters with differing relevant subspaces. This generation method thus takes into account that objects can be relevant for several clusters in multiple views.

Furthermore we demonstrate the performance of ASCLU on two real world data sets (iris and pendigits) provided by the UCI repository [7]. Since the motivating assumption for alternative clustering is the presence of multiple views in the data, traditional class-based evaluation, where each object is assigned to exactly one class, is not reasonable in this case. We therefore examine clustering results for the pendigits data set visually, similar to [4]. For the iris data set we use the method in [9] and concatenate the original data with random permutations of itself, which results in one high dimensional data set containing several views, the concatenations. For the quality assessment the F1-value is qualified, as it can handle overlapping clusters and classes and is used for evaluation of subspace clusters [2, 13, 9].

Experiments on synthetic data.

The first experiment serves to examine the ability of ASCLU to calculate a real alternative Res to a given clustering $Known$. An alternative clustering should yield new information compared to the given clustering. This can be characterized by clusters that either group similar objects compared to the given clusters but in different subspaces or simply group other objects than the given ones. A small synthetic data set with 300 objects and 16 dimensions, where 16 clusters are hidden in five different subspaces, allows an easy visual examination. Fig. 2 depicts a representation of the data matrix, where the given clusters (black boxes) and all clusters found by ASCLU (red circles) are plotted. Each column of the matrix represents a database object, each row represents a dimension. For a clear presentation, the objects and dimensions have been permuted, such that the given clusters and several categories of new information types become apparent. The black rectangular area represents the previously known information. New clusters should preferably avoid this area of given clusters and concentrate on new information. The black area is only sparsely populated with circles, which indicates that the newly found clusters do not provide the same information. To gain new knowledge a cluster has to cover a sufficient amount of new objects or/and different subspaces. Fig. 2 also shows that ASCLU does not block the given cluster area completely for the new clustering, like other approaches do for dimensions [17], but allows for clusters to overlap regarding dimensions or objects of the given clusters. The potential information content of the new clustering is thus extended.

Experiments on UCI data sets.

In the following we focus on the effectiveness of ASCLU for real world data sets. We start using the pendigits data set to show that ASCLU reveals new patterns compared to some given ones. The pendigits dataset is very rewarding for clustering analysis, since cluster results are very descriptive and visualizable. This dataset is thus suitable to examine, whether ASCLU is able to find valuable alternatives for given subspace clusters in real world data. As input for ASCLU we use three clusters: digit 0 and digit 6, each clustered in the first 3 xy-coordinates, and digit 9 clustered in the last 3 ones. Given this input, ASCLU determines subspace clusters with deviating properties and thus novel information. The first indicator is that the given three digits appear in less clusters than the other digits, which shows, that ASCLU is more interested in the unknown (not given)

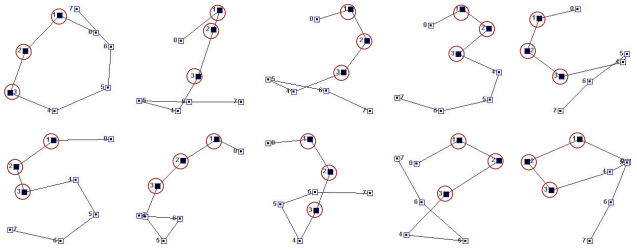

Figure 3: Alternative subspace cluster for pendigits

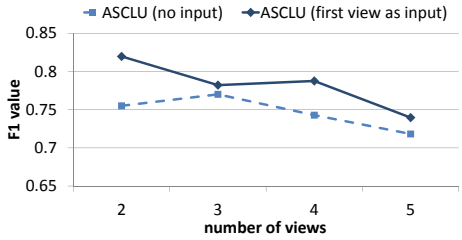

Figure 4: Quality of ASCLU on the iris dataset

digits. A second observation is that the given digits appear mainly in clusters that cover nearly all digits, thus representing novel object groupings. In Fig. 3 such an identified cluster is illustrated, where all digits have similar values for the marked y-coordinates. Only in very few clusters the three digits are again clustered individually; though, these clusters yield novel attribute information of the digits.

The next experiment on the iris data evaluates the influence of the input clustering on the quality of ASCLU's results. As described, we extend the iris data set to multiple views per object. This way we are able to determine a subspace clustering as input, which will in this case be the first view. We therefore only consider the classes of the other views, the latter concatenations, as ground truth to compare the results with. In Fig. 4 we compare the results of ASCLU with and without this first view as input clustering for three different data sets, which differ in the number of concatenated views. The results show, that a high quality input clustering, like one view in this example, has a positive effect on the quality of the alternatively found clusters in the other views. This effect is explainable by the fact that due to the given clusters ASCLU already excludes several clusters with similar informations to avoid redundancy. These avoided similar clusters do obviously not belong to the ground truth and do often lead to the redundancy and thus the exclusion of valuable clusters in other views. As Fig. 4 shows this effect is best traceable if there is only one view besides the given one, but the given information has also a positive effect on more than just one view.

5. CONCLUSION

In this paper, we propose a method for detecting alternative subspace clusterings. In contrast to previous approaches that determine alternative groupings, we specifically consider the relevant dimensions of each subspace cluster to identify different views within the data. Besides generating deviating clusters compared to the given input clustering, our model ensures that each resulting cluster provides novel knowledge: our model avoids generating redundant clusters. The experimental evaluation confirms that our model successfully detects meaningful alternative subspace clusters based on the given input clustering.

Acknowledgment

This work has been supported by the UMIC Research Centre, RWTH Aachen University, Germany.

6. REFERENCES

[1] 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.

[2] 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.

[3] E. Bae and J. Bailey. Coala: A novel approach for the extraction of an alternate clustering of high quality and high dissimilarity. In *ICDM*, pages 53–62, 2006.

[4] Y. Cui, X. Z. Fern, and J. G. Dy. Non-redundant multi-view clustering via orthogonalization. In *ICDM*, pages 133–142, 2007.

[5] I. Davidson and Z. Qi. Finding alternative clusterings using constraints. In *ICDM*, pages 773–778, 2008.

[6] 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.

[7] A. Frank and A. Asuncion. UCI machine learning repository, 2010. http://archive.ics.uci.edu/ml.

[8] D. Gondek and T. Hofmann. Non-redundant clustering with conditional ensembles. In *KDD*, pages 70–77, 2005.

[9] 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.

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

[11] M. Kamber and J. Han. *Data mining: Concepts and techniques*. Morgan Kaufmann, San Francisco, 2001.

[12] 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.

[13] G. Moise and J. Sander. Finding non-redundant, statistically significant regions in high dimensional data: a novel approach to projected and subspace clustering. In *KDD*, pages 533–541, 2008.

[14] 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.

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

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

[17] Z. Qi and I. Davidson. A principled and flexible framework for finding alternative clusterings. In *KDD*, pages 717–726, 2009.