Proc. IEEE International Conference on Data Mining (ICDM 2007), Omaha, Nebraska, USA, pages 409-414

DUSC: Dimensionality Unbiased Subspace Clustering

Ira Assent Ralph Krieger Emmanuel Müller Thomas Seidl
Data management and data exploration group, RWTH Aachen University, Germany
{assent,krieger,mueller,seidl}@cs.rwth-aachen.de

Abstract

To gain insight into today's large data resources, data mining provides automatic aggregation techniques. Clustering aims at grouping data such that objects within groups are similar while objects in different groups are dissimilar. In scenarios with many attributes or with noise, clusters are often hidden in subspaces of the data and do not show up in the full dimensional space. For these applications, subspace clustering methods aim at detecting clusters in any subspace. Existing subspace clustering approaches fall prey to an effect we call dimensionality bias. As dimensionality of subspaces varies, approaches which do not take this effect into account fail to separate clusters from noise. We give a formal definition of dimensionality bias and analyze consequences for subspace clustering. A dimensionality unbiased subspace clustering (DUSC) definition based on statistical foundations is proposed. In thorough experiments on synthetic and real world data, we show that our approach outperforms existing subspace clustering algorithms.

1 Introduction

Increasingly large data resources in life sciences, mobile information and communication, e-commerce, and other application domains require automatic techniques for gaining knowledge. One of the major knowledge discovery tasks is clustering. It aims at summarizing data base objects such that similar objects are grouped together while dissimilar ones are separated. In noisy data or data with many attributes, clusters are often hidden in subspaces of the attributes and do not show up across the full attribute space. A global reduction to relevant attributes is often infeasible, as relevance of attributes is not necessarily globally uniform. Varying relevance of attributes for individual clusters requires clustering over any possible subset of the attributes. Subspace clustering therefore aims at detecting clusters in any possible attribute combination. For traditional "full-space" clustering, different paradigms exist. Density-based approaches have shown to successfully mine clusters even in the presence of noise. The idea is to define clusters as dense areas separated by sparsely populated areas. Density of an object is measured either by mere counting of objects or by more complex functions on the number and location of objects in the neighborhood. An object is considered dense if its density is above some threshold. Density-based clustering has been extended to subspace clustering in previous works. The definition of density is typically similar to that in full space clustering. However, ignoring the dimensionality in subspace clustering has serious consequences for the quality of the result. Density in subspaces of different dimensionalities is not comparable. Existing approaches which do not take this effect into consideration hence check incomparable values against the same threshold. Thus, they fail to separate dense from sparse regions across subspaces of different dimensionalities. Assuming a simple setup of uniformly distributed data, we show that density measures which ignore dimensionality cannot distinguish this pseudo-cluster scenario from true clusters in all subspaces. As a consequence, dimensionality bias means failing at the very core of density-based subspace clustering. Hence, existing subspace clustering algorithms either lose clusters or detect numerous pseudo-clusters depending on the parameter setting.

In this work, we focus on eliminating dimensionality bias. We propose a new density-based subspace clustering approach DUSC (dimensionality unbiased subspace clustering) based on statistical foundations which takes the dimensionality into account. We show that this method eliminates dimensionality bias and leads to comparable clustering results between subspaces of different dimensionalities. To ensure efficient mining of density-based subspace clusters, we derive powerful pruning properties.

Summing up, our contributions include:

- definition and analysis of dimensionality bias and its consequences for subspace clustering
- definition of density based on statistical foundations
- dimensionality unbiased subspace clustering model
- powerful pruning properties

2 Related Work

In density-based clustering, clusters are dense areas separated by sparsely populated areas as in DBSCAN [4]. These methods have shown to be capable of detecting arbitrarily shaped clusters even in noisy settings. Neighborhoods are checked for a minimum number of points. The distribution of objects is ignored and sensitivity to parameter settings is a challenge. Traditional clustering algorithms do not scale to multi-dimensional or high-dimensional spaces. As clusters do not show across all attributes, they are hidden by irrelevant attributes [3]. Subspace clustering aims at mining clusters in arbitrary, possibly overlapping, subspace projections [12]. CLIQUE discretizes the data using grids and uses monotonicity on density of cells for pruning [1]. Grids greatly reduce the computational complexity, yet clusters which spread across several cells might be missed. MAFIA extends CLIQUE to a data-adapted grid to reduce the number of clusters lost [10]. SCHISM extends CLIQUE using a variable threshold to cope with different dimensionalities, yet relies on heuristics and a grid-based discretization for pruning [13]. Consequently, completeness is lost as in all grid-based approaches. Specialized algorithms for categorical data or sequences [15, 2] require discretization as well. SUBCLU uses a density monotonicity property to prune subspaces [6]. As dimensionality is ignored, it suffers from dimensionality bias, i.e. clusters cannot be separated from noise across subspaces. FIRES is a generic framework which relies on approximative techniques in a filter-refinement scheme [9].

3 Density-Based Subspace Clusters

In many applications clusters are hidden in subspaces and cannot be revealed by any cluster analysis that mines all dimensions simultaneously. Subspace clustering automatically focuses to the respectively relevant dimensions.

Let $\mathbf{U} = [0, \mathbf{v}]$ be a universal domain for all dimensions, $\mathbf{D} = \{1, \ldots, d\}$ be an index set, and $\mathbf{DB} \subseteq \mathbf{U^D}$ a d-dimensional database with n objects. A subspace $\mathbf{U^S}$ is the projection of $\mathbf{U^D}$ to the r dimensions specified by the index set $\mathbf{S} = \{s_1, \ldots, s_r\} \subseteq \mathbf{D}$. Analogously, let $\mathbf{DB^D}$ denote the original database and $\mathbf{DB^S}$ its projection to the dimensions in $\mathbf{S}$. For ease of notation, we refer to a subspace $\mathbf{U^S}$ by its index set $\mathbf{S}$. The definition of density-based subspace clusters extends standard notions in density-based clustering [4]. Let $||.||^{\mathbf{S}}$ denote the restriction of norm $||.|| : \mathbf{U^D} \rightarrow \mathcal{R}$ to the dimensions in subspace $\mathbf{S}$. The area of influence is the neighborhood in subspace $\mathbf{S}$:
$\mathcal{N}_\varepsilon^{\mathbf{S}}(o) = \{p | p \in \mathbf{DB}, \|p - o\|^{\mathbf{S}} \leq \varepsilon\}$. Typically, density of an object o is determined by simply counting the number of objects in a fixed ε-range $\mathcal{N}_\varepsilon^{\mathbf{S}}(o)$. We generalize this idea by assigning weights to each object contained in $\mathcal{N}_\varepsilon^{\mathbf{S}}(o)$.

Definition 1 ***Density Measure***
Let $\mathcal{W}$ be an arbitrary weighting function $\mathcal{W} : \mathcal{R} \rightarrow \mathcal{R}$. Based on $\mathcal{W}$, a generalized density measure $\varphi^{\mathbf{S}}(o)$ for an object o in subspace $\mathbf{S}$ is defined as:

$$\varphi^{\mathbf{S}}(o) = \sum_{p \in \mathcal{N}_\varepsilon^{\mathbf{S}}(o)} \mathcal{W}\left(\|o - p\|^{\mathbf{S}}\right)$$

Thus, an object o in subspace $\mathbf{S}$ is called *dense* if the weighted distances to objects in its area of influence sum up to more than a given density threshold $\varphi^{\mathbf{S}}(o) \geq \tau$.

3.1 Dimensionality Bias

Subspace clustering methods analyze data spaces of different dimensionalities. Therefore, avoiding an effect which we call dimensionality bias is an important issue. Dimensionality bias refers to a dependency of density on the dimensionality of the subspace: as dimensionality increases, average distances between objects increase and cluster radii grow. At the same time, the expected density within the area of influence drops accordingly. Thus, ignoring the dependency of density on the dimensionality of the subspace leads to incomparable density values. Incomparable density values pose the following problem: the high discrepancy in density scales of low-dimensional or high-dimensional subspaces makes it impossible to find a suitable parameter for a fixed density threshold τ. If on the one hand τ is parametrized such that high-dimensional clusters with low expected density are detected then numerous excess pseudo-clusters are generated in low-dimensional spaces where expected density is high. On the other hand, a parametrization of τ which separates clusters from noise in low-dimensional spaces loses clusters in high-dimensional spaces. We assume that τ is fixed as dimensionality dependent thresholds can also be incorporated into the density measure (the same argument holds if one were to vary ε).

To obtain comparable density values, unbiased density measures have to be independent of the dimensionality of the subspace. Statistically speaking, this corresponds to the same expected density value regardless of the dimensionality of the subspace.

Definition 2 ***Dimensionality Unbiased Density Measure***
A density measure $\varphi^{\mathbf{S}}$ is dimensionality unbiased if its expected density is the same for any two subspaces $\mathbf{S}_1$ and $\mathbf{S}_2 \subseteq \mathbf{D}$:

$$\forall\, \mathbf{S}_1, \mathbf{S}_2 : \;\; E\left[\varphi^{\mathbf{S_1}}\right] = E\left[\varphi^{\mathbf{S_2}}\right]$$

We now show how dimensionality bias can be eliminated for any density estimator. As the expected density should be the same for any two subspaces, we normalize density estimators with their expected density. For any density measure $\varphi^{\mathbf{S}}$, the normalized measure $\frac{1}{E[\varphi^{\mathbf{S}}]}\varphi^{\mathbf{S}}$ is dimensionality unbiased. With linearity property of the expectation, this is straightforward: $E[\frac{1}{E[\varphi^{\mathbf{S}}]}\varphi^{\mathbf{S}}] = \frac{1}{E[\varphi^{\mathbf{S}}]}E[\varphi^{\mathbf{S}}] = 1$ for all subspaces. Thus, for any two subspaces, normalizing the density measure by the expected value of the subspace yields comparable density values for any two subspaces $\mathbf{S}_1$ and $\mathbf{S}_2$. Normalization could be achieved by other means such as subtracting the expected value, but, as we will see later, dividing by the expected value simplifies the choice of density parameters in subspace clustering.

3.2 An Unbiased Density Estimator

In this section we use statistical analysis to develop an unbiased density measure for subspace clustering. In statistics, $kernel\ estimators$ are used to estimate density functions from a set of data objects. A kernel weights the observations in the data set to compute the density value at any position in the data space.

Using a kernel function which assigns higher values to closer objects and lower values to objects further away, density is more accurately measured than by mere counting of objects within the neighborhood [5, 14]. The most commonly used are Gauss, Epanechnikov, Bisquare and Triangular kernels. Gauss, however, assigns non-zero values to all objects in the data base, even to those at very large distances. It is a poor density estimator in terms of efficiency and effectivity [14]. The Epanechnikov kernel is both an efficient and effective choice, since it is computationally efficient and minimizes the mean integrated squared error [14]. Thus, we use Epanechnikov kernel in the following, but in principle any other kernel could be used as well.

Within an area of influence, the Epanechnikov kernel assigns decreasing weights to objects with increasing distance. For a subspace $\mathbf{S}$, the Epanechnikov kernel function $K^{\mathbf{S}}$ is defined as:

$$K^{\mathbf{S}}(x) = \begin{cases} \frac{|\mathbf{S}|+2}{2c_{|\mathbf{S}|}}\left(1 - \left(\|x\|^{S}\right)^2\right), & \|x\|^{\mathbf{S}} \leq 1 \\ 0, & \text{else.} \end{cases}$$

where $|\mathbf{S}|$ denotes the dimensionality of the subspace and $c_{|\mathbf{S}|} = \frac{\pi^{|\mathbf{S}|/2}}{\Gamma(|\mathbf{S}|/2+1)}$ is the volume of the $|\mathbf{S}|$-dimensional unit sphere; gamma function $\Gamma(n+1) = n * \Gamma(n), \Gamma(1) = 1, \Gamma(1/2) = \sqrt{\pi}$. Each kernel is scaled in width according to a *bandwidth* ε which corresponds to the area of influence of a density based subspace clustering algorithm. For subspace clustering, we need only the Epanechnikov kernel weights $1 - \left(\|x\|^{S}\right)^2$ to obtain the following weighting function according to Definition 1 :

Definition 3 ***Epanechnikov Density Measure***
Let $\mathcal{W}(t) = 1 - t^2$ be the Epanechnikov weighting function. We define the Epanechnikov density measure for an area of influence specified by ε as:

$$\varphi^{\mathbf{S}}(o) = \sum_{p \in \mathcal{N}_{\varepsilon}^{\mathbf{S}}(o)} \left(1 - \left(\frac{\|o-p\|^{\mathbf{S}}}{\varepsilon}\right)^2\right)$$

As seen above, we can remove dimensionality bias by taking the expected density for subspaces into account. As clustering aims at detecting dense regions in a given data set, clusters should have higher density values than data without any clusters. A data set without clusters corresponds to uniformly distributed data, i.e. all values are taken with the same probability. By requiring that density should exceed the expected density of uniformly distributed subspaces, we ensure that no pseudo-clusters are "detected". By applying this on the Epanechnikov density measure $\varphi^{\mathbf{S}}$ we obtain the unbiased Epanechnikov density measure:

Definition 4 ***Unbiased Epanechnikov Density Measure***
The unbiased density measure for the Epanechnikov influence function $\varphi^{\mathbf{S}}$ is given by

$$\frac{1}{\alpha(\mathbf{S})}\varphi^{\mathbf{S}}(o) \ \ with$$

$$\alpha(\mathbf{S}) = E_{\mathbf{S}}\left[\varphi^{\mathbf{S}}(o)\right] = \frac{2n\varepsilon^{|\mathbf{S}|}c_{|\mathbf{S}|}}{\mathbf{v}^{|\mathbf{S}|}(|\mathbf{S}|+2)} \qquad (1)$$

Dimensionality bias can be removed for other kernel density estimators as well by normalizing the density measure with the reciprocal expected density. The statistical approach has two advantages: the effectiveness of kernel estimators has been studied in theoretical and practical settings, and computation of the expected density for probability density functions follows standard methods.

4 DUSC Subspace Clustering

Intuitive density threshold. The density threshold is a core parameter since it sets the dividing line between dense objects and noise. As this parameter has to be set by the user it is important for users to have an intuitive understanding of this parameter. Commonly, users do not know density distribution apriori, which makes the choice of a density value difficult. We exploit the fact that in our approach density is measured with respect to the expected density as discussed before. Consequently, users do not need to specify absolute density thresholds, but only a factor by which the expected density has to be exceeded. Following the definition in the previous section, an object o is dense in subspace $\mathbf{S}$ according to the expected density $\alpha(\mathbf{S})$ iff:

$$\frac{1}{\alpha(\mathbf{S})}\varphi^{\mathbf{S}}(o) \geq \mathbf{F}$$

where $\mathbf{F}$ denotes the density threshold. As the density factor $\mathbf{F}$ is independent of the dimensionality and data set size, it is much easier to specify than traditional density thresholds. Moreover, we demonstrate in the experiments that this parameter is robust with a setting of $\mathbf{F} > 50$ for many applications.

Empty space problem. With increasing dimensionality the expected density and hence the expected number of objects contained in an area of influence drops exponentially [3]. This effect is termed "empty space problem" in statistics [14]. Compared to the expected density an object may be determined as dense even if the area of influence is nearly devoid of observations, resulting in pseudo-dense single objects. To remove pseudo-dense objects we introduce a specific density constraint on the expected density of η objects in the area of influence. The expected density value of an object o which contains η objects in the area of influence $E_\eta\left[\frac{1}{\alpha(\mathbf{S})}\varphi^{\mathbf{S}}(o)\right]$ can be derived as follows:

$$\begin{aligned}\frac{1}{\alpha(\mathbf{S})}\varphi^{\mathbf{S}}(o) &\geq E_\eta\left[\frac{1}{\alpha(\mathbf{S})}\varphi^{\mathbf{S}}(o)\right]\\ \varphi^{\mathbf{S}}(o) &\geq \eta\frac{2}{|S|+2}\\ \varphi^{\mathbf{S}}(o) &\geq \eta\cdot\omega(\mathbf{S}) \qquad \left[\omega(\mathbf{S}) := \frac{2}{|S|+2}\right] \quad (2)\end{aligned}$$

To guarantee that objects are not considered dense if the ε sphere is virtually empty, a very small value for η is sufficient (generally two or three). Users typically do not need to change this value. Our new density based subspace clustering model below combines the density constraints α and ω given in formulae (1) and (2). α and ω combined ensure an unbiased density notion without defining objects in nearly empty regions as dense.

Redundancy. Since the number of possible subspace projections is exponential in the number of dimensions, subspace clustering algorithms often produce numerous redundant subspace clusters. To avoid excessive cluster outputs which contain essentially the same information repeated in different dimensionalities, we check if a cluster $\mathbf{C}$ in subspace $\mathbf{S}$ is redundant. We define a cluster as redundant if (most of) the objects contained in the cluster are also contained in another cluster in a higher dimensional subspace $\mathbf{S}' \supset \mathbf{S}$. We use a parameter r to specify the degree of redundancy acceptable to the user. To restrict the output to a reasonable size a strict redundancy parameter is often appropriate ($r \approx 0.1$).

So far, we have studied the density of individual objects. Subspace clusters, following density-based clustering paradigm, are connected sets of such dense objects. To ensure that clusters reflect the inherent structure of the data, they should contain a certain minimum number of objects. This constraint $minSize$ is typically about 1% of the database size.

The resulting subspace cluster model taking these conclusions into account is formalized in the following.

Definition 5 ***DUSC Subspace Cluster***
A set of objects $\mathbf{C} \subseteq \mathbf{DB}$ in subspace $\mathbf{S} \subseteq \mathbf{D}$ is a subspace cluster if:

- *objects in $\mathbf{C}$ are **$\mathbf{S}$-connected:***
$\forall o, p \in \mathbf{C} : \exists k : \forall i = 1, \ldots, k-1 : \exists\, q_i \in \mathbf{C} : \|q_i - q_{i+1}\|^{\mathbf{S}} \leq \varepsilon \wedge q_1 = o, q_k = p$
- ***more dense than expected and not pseudo-dense:***
$\forall o \in \mathbf{C}$*:* $\varphi^{\mathbf{S}}(o) \geq max\{\mathbf{F}\cdot\alpha(\mathbf{S}), \eta\cdot\omega(\mathbf{S})\}$
- $\mathbf{C}$ *is **maximal**, i.e. contains all S-connected objects*
$\forall o, p \in \mathbf{DB}$*,* o, p *S-connected* $\Rightarrow (o \in \mathbf{C} \Leftrightarrow p \in \mathbf{C})$
- ***minimum cluster size:*** $|\mathbf{C}| \geq minSize$
- ***not redundant:*** $\neg\exists(\mathbf{C}', \mathbf{S}')$ *subspace cluster with* $\mathbf{C}' \subseteq \mathbf{C} \wedge \mathbf{S} \subset \mathbf{S}' \wedge |\mathbf{C}'| \geq r \cdot |\mathbf{C}|$

The DUSC subspace clustering model extends existing density-based notions of maximality and connectedness with statistically sound density computation via normalized Epanechnikov kernel and expected density. Clusters contain a significant part of the data, and are not redundant.

5 Efficient subspace clustering

As subspace clustering mines clusters in multi-dimensional and high-dimensional data spaces, evaluating the cluster model in a naive way is infeasible as the number of possible subspaces (and subspace clusters) is exponential with the dimensionality. We propose an efficient algorithm which combines three paradigms for improving the runtime:

(1) a filter-and-refine architecture with a filter step based on weak density monotonicity for pruning the search space

(2) a depth first approach which avoids excess candidate generation on a specialized index structure

(3) redundancy pruning: mine lower dimensional projections only if no redundant higher dimensional cluster was found in this region during depth first search

Pruning requires a monotonicity on some property of subspace clusters. A region which does not form a subspace cluster in some dimensionality, implies that this region cannot be a subspace cluster in any higher dimensional subspace, and we may safely prune this region from further consideration. As the density definition given in Definition 5 depends on the dimensionality of the analyzed subspace, it is not monotonous in the above sense and thus cannot be used directly for pruning. The higher-dimensional a subspace, the lower its expected density. Thus, a region which is not dense according to a low-dimensional subspace's density threshold, may be dense with respect to a higher-dimensional subspace's threshold. To overcome this problem, we introduce the new concept of weak density threshold.

Definition 6 ***Weak density***
An object o in a subspace $\mathbf{S}$ is defined as ***weak dense*** *if:*

$$\varphi^{\mathbf{S}}(o) \geq max\{\mathbf{F} \cdot \alpha(\mathbf{D}), \eta \cdot \omega(\mathbf{D})\}$$

The weak density definition uses the highest, $|\mathbf{D}|$-dimensional threshold $max\{\mathbf{F} \cdot \alpha(\mathbf{D}), \eta \cdot \omega(\mathbf{D})\}$ for the density of an object o in subspace $\mathbf{S}$. Thus if an object o is not weakDense, o cannot be dense in any super subspace of $\mathbf{S}$. Moreover, density-connected sets can be pruned if they contain less than $minSize$ objects. Pruning based on these two properties, called *weak monotonicity*, is used by the DUSC algorithm to efficiently prune the search space in a depth first search. It is valid in the sense that no cluster is wrongfully dropped from consideration. Due to space limitations we defer proofs to an extended version of this paper.

6 Experiments

We ran extensive experiments on real world data (Pendigits, Glass, Vowel [11] and Shapes [8]) to demonstrate the accuracy of the DUSC subspace clustering model. Synthetic data was used to demonstrate the efficiency and scalability for large and high-dimensional data sets and to validate that indeed all subspace clusters hidden in the data are found. Experiments were run on Pentium 4 machines with 2.4 Ghz and 1 GB main memory.

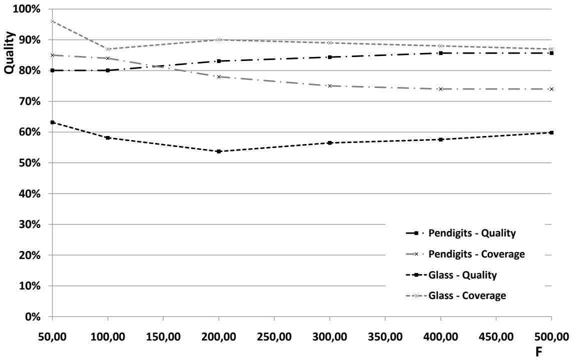

Figure 1. Quality vs. density threshold F

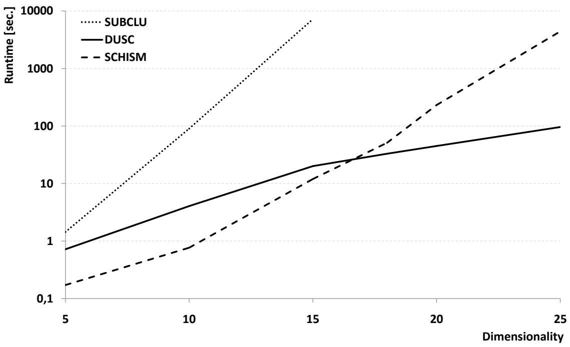

Figure 2. Scalability wrt. dimensionality

	DUSC 0%		DUSC 5%		DUSC 10%		SUBCLU		SCHISM	
	Q	C	Q	C	Q	C	Q	C	Q	C
Pendigits	86	74	83	87	81	92	58	100	77	100
Glass	60	87	51	90	50	93	44	100	44	99
Vowel	82	70	79	100	74	100	10	100	42	100
Shape	100	31	100	31	100	31	98	82	100	1

Figure 3. Accuracy for real data sets

Accuracy of the subspace clustering is determined in terms of quality and coverage. Corresponding roughly to the measures of precision and recall, quality accounts for purity of the clustering, while coverage measures the size of the clustering. Quality is determined using the entropy, i.e. $H(\mathbf{C}) = -\sum_{i=1}^{k} p(i|\mathbf{C}) \cdot \log(p(i|\mathbf{C}))$ for k class labels in cluster $\mathbf{C}$. For a set of clusters we take the average entropy weighted by the number of objects per cluster. For readability, we take the inverse entropy and normalize it to a range of 0% to 100% by dividing by the maximum entropy $(1 - H(\mathbf{C})/log(k))$. Coverage is the percentage of objects in any subspace cluster. It indicates the ratio of clustered objects to noise. The amount of noise in a data set is typically not known apriori, but noise is present in most real world data sets. As sparsely populated regions often exhibit a weak correlation to the class label, quality can be improved if less objects belong to a cluster (coverage drops). Thus we always evaluate quality and coverage in combination. Our algorithm has few and intuitive parameters. They can be easily understood by users and are very robust on different data sets. Recall that $minSize$ of a cluster is the minimum number of objects per subspace cluster. Values around 1% of the data lead to manageable result sizes. Lower values produce more subspace clusters than users might want to study, whereas higher values diminishes the output size. η is fixed to two to eliminate the empty-space problem. Obviously, at least two objects should be contained even in very high dimensional subspace clusters. The density-thresholds α and ω are directly computed from the expected density. Users only provide a factor $\mathbf{F}$ which describes by how much the expected density should be exceeded. This $\mathbf{F}$ is independent of the data base size and its dimensionality. Finally, the bandwidth ε regulates the kernel density. This parameter can be estimated using standard methods from statistics

[14]. We demonstrate robustness of our DUSC algorithm on two real world data sets with very different data distributions: Pendigits and Glass. The first experiment studies the effect of density threshold $\mathbf{F}$ on quality and coverage. The results shown in Figure 1 confirm that DUSC is remarkably robust with respect to $\mathbf{F}$. To evaluate the performance of DUSC we generated synthetic data of different dimensionalities and hid clusters (some overlapping) in different subspaces. Moreover, varying densities for individual clusters and different numbers of dense objects per cluster have been included. Additionally, noise, i.e. objects which do not belong to any subspace clusters, has been added. The results in Figure 2 show that the DUSC algorithm efficiently mines subspace clusters and clearly scales better than competing algorithms. We evaluated the accuracy [quality (Q) and the coverage (C)] of DUSC against SUBCLU and SCHISM using real world data sets. For DUSC we used the default values for $\mathbf{F}$ according to the heuristic in section 4. As SUBCLU and DUSC are both density based clustering algorithms we used the heuristic presented in [7] to determine ε. SCHISM is a grid based approach. Its parameters ξ, τ are also determined using the original heuristic in [13]. For their third parameter u which cuts off low dimensional clusters, we noticed that the heuristic does not yield good results in terms of accuracy. To obtain better quality results for SCHISM that explain more about the true differences between different density models, we used an even lower value for u than suggested by the authors. However as SCHISM is a grid based approach it still does not reach the quality of DUSC. The first column in Figure 3 shows the quality results of DUSC with redundancy set to zero which are the best measured qualities for all data sets. Allowing more redundancy, coverage increases significantly and quality goes only slightly down. However, even for $r = 10\%$ DUSC shows better quality than the competing algorithms. The fact that coverage is not 100% indicates that DUSC can distinguish between noise and clusters in subspaces of varying dimensionalities. The pendigits data set, for example, contains handwritten numbers, some of which are clearly different from the rest of the data set. Biased algorithms like SCHISM and SUBCLU do not detect noise, but assign all objects to clusters. The last data set SHAPE contains rotated versions of 9 different shapes, but only 3 of the shapes clearly form clusters. Thus most of the objects have to be considered noise. DUSC detects the given clusters correctly while SCHISM detects only a small part of the clusters and SUBCLU mixes up clusters with noise (less than 100% quality).

7 Conclusion

We introduced DUSC, an efficient density-based subspace clustering algorithm. Using both statistical density estimation and expected density in varying subspaces, we are capable of accurately grasping the inherent data structure without dimensionality bias. Our experiments on large high-dimensional synthetic and real world data sets show that DUSC outperforms other subspace clustering algorithms in terms of accuracy and runtimes.

Acknowledgments: This research was funded in part by the cluster of excellence on Ultra-high speed Mobile Information and Communication (UMIC) of the DFG (German Research Foundation grant EXC 89).

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, B. Glavic, and T. Seidl. Spatial multidimensional sequence clustering. In *SSTDM at ICDM*, 2006.

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

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

[5] A. Hinneburg and D. Keim. An efficient approach to clustering in large multimedia databases with noise. In *KDD*, pages 58–65, 1998.

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

[7] K. Kailing, H.-P. Kriegel, P. Kröger, and S. Wanka. Ranking interesting subspaces for clustering high dimensional data. In *PKDD*, pages 241–252, 2003.

[8] E. Keogh, L. Wei, X. Xi, S.-H. Lee, and M. Vlachos. LB_Keogh supports exact indexing of shapes under rotation invariance with arbitrary representations and distance measures. In *VLDB*, pages 882–893, 2006.

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

[10] H. Nagesh, S. Goil, and A. Choudhary. MAFIA: Efficient and scalable subspace clustering for very large data sets. In *TR 9906-010, NWU*, 1999.

[11] D. Newman, S. Hettich, C. Blake, and C. Merz. UCI repository of MLDBs, 1998.

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

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

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

[15] M. Zaki, M. Peters, I. Assent, and T. Seidl. CLICKS: An effective algorithm for mining subspace clusters in categorical datasets. *DKE*, 60:51–70, January 2007.