Data Mining and Knowledge Discovery, 11, 5–33, 2005
© 2005 Springer Science + Business Media, Inc. Manufactured in The Netherlands.

# Automatic Subspace Clustering of High Dimensional Data

RAKESH AGRAWAL ragrawal@almaden.ibm.com
*IBM Almaden Research Center, 650 Harry Road, San Jose, CA 95120*

JOHANNES GEHRKE* johannes@cs.cornell.edu
*Computer Science Department, Cornell University, Ithaca, NY*

DIMITRIOS GUNOPULOS* dg@cs.ucr.edu
*Department of Computer Science and Eng., University of California Riverside, Riverside, CA, 92521*

PRABHAKAR RAGHAVAN* pragh@verity.com
*Verity, Inc*

**Editor:** Geoff Webb

**Abstract.** Data mining applications place special requirements on clustering algorithms including: the ability to find clusters embedded in subspaces of high dimensional data, scalability, end-user comprehensibility of the results, non-presumption of any canonical data distribution, and insensitivity to the order of input records. We present CLIQUE, a clustering algorithm that satisfies each of these requirements. CLIQUE identifies dense clusters in subspaces of maximum dimensionality. It generates cluster descriptions in the form of DNF expressions that are minimized for ease of comprehension. It produces identical results irrespective of the order in which input records are presented and does not presume any specific mathematical form for data distribution. Through experiments, we show that CLIQUE efficiently finds accurate clusters in large high dimensional datasets.

**Keywords:** subspace clustering, clustering, dimensionality reduction

## 1. Introduction

Clustering is a descriptive task that seeks to identify homogeneous groups of objects based on the values of their attributes (dimensions) (Jain and Dubes, 1988; Kaufman and Rousseeuw, 1990). Clustering techniques have been studied extensively in statistics (Arabie and Hubert, 1996), pattern recognition (Duda and Hart, 1973; Fukunaga, 1990), and machine learning (Cheeseman and Stutz, 1996; Michalski and Stepp, 1983). Recent work in the database community includes CLARANS (Ng and Han, 1994), Focused CLARANS (Ester et al., 1995), BIRCH (Zhang et al., 1996), DBSCAN (Ester et al., 1996) and CURE (Guha et al., 1998).

*Work done while the author was at IBM Almaden.

Current clustering techniques can be broadly classified into two categories (Jain and Dubes, 1988; Kaufman and Rousseeuw, 1990): *partitional* and *hierarchical*. Given a set of objects and a clustering criterion (Sneath and Sokal, 1973), partitional clustering obtains a partition of the objects into clusters such that the objects in a cluster are more similar to each other than to objects in different clusters. The popular $K$-means and $K$-medoid methods determine $K$ cluster representatives and assign each object to the cluster with its representative closest to the object such that the sum of the distances squared between the objects and their representatives is minimized. CLARANS (Ng and Han, 1994), Focused CLARANS (Ester et al., 1995), and BIRCH (Zhang et al., 1996) can be viewed as extensions of this approach to work against large databases. Mode-seeking clustering methods identify clusters by searching for regions in the data space in which the object density is large. DBSCAN (Ester et al., 1996) finds dense regions that are separated by low density regions and clusters together the objects in the same dense region.

A hierarchical clustering is a nested sequence of partitions. An agglomerative, hierarchical clustering starts by placing each object in its own cluster and then merges these atomic clusters into larger and larger clusters until all objects are in a single cluster. Divisive, hierarchical clustering reverses the process by starting with all objects in cluster and subdividing into smaller pieces (Jain and Dubes, 1988). CURE (Guha et al., 1998) is an extension of this approach that scales well with the size of the dataset.

### 1.1. *Desiderata from the data mining perspective*

Emerging data mining applications place the following special requirements on clustering techniques, motivating the need for developing new algorithms.

***Effective treatment of high dimensionality.*** An object (data record) typically has dozens of attributes and the domain for each attribute can be large. It is not meaningful to look for clusters in such a high dimensional space as the average density of points anywhere in the data space is likely to be quite low (Berchtold et al., 1997). Compounding this problem, many dimensions or combinations of dimensions can have noise or values that are uniformly distributed. Therefore, distance functions that use all the dimensions of the data may be ineffective. Moreover, several clusters may exist in different subspaces comprised of different combinations of attributes.

***Interpretability of results.*** Data mining applications typically require cluster descriptions that can be easily assimilated by an end-user as insight and explanations are of critical importance (Fayyad et al., 1996). It is particularly important to have simple representations because most visualization techniques do not work well in high dimensional spaces. However, the description of the clusters should be accurate enough that detail is not lost. This is a common failing of K-means type techniques: since the distance metric used is typically the Manhattan or the Euclidian distance, the clusters found can only have near spherical shape (Guha et al., 1998).

***Scalability and usability.*** The clustering technique should be fast and scale with the number of dimensions and the size of input. It should be insensitive to the order in which the data records are presented. It should not presume some canonical form for the data distribution.

Current clustering techniques do not address all these points adequately, although considerable work has been done in addressing each point separately.

The problem of high dimensionality is often tackled by requiring the user to specify the subspace (a subset of the dimensions) for cluster analysis (e.g. (Internationl Business Machines, 1996)). However, user-identification of subspaces is quite error-prone. Another way to address high dimensionality is to apply a dimensionality reduction method to the dataset. Methods such as principal component analysis or Karhunen-Loève transformation (Duda and Hart, 1973; Fukunaga, 1990) optimally transform the original data space into a lower dimensional space by forming dimensions that are linear combinations of given attributes. The new space has the property that distances between points remain approximately the same as before. While these techniques may succeed in reducing the dimensionality, they have two shortcomings. First, the new dimensions can be difficult to interpret, making it hard to understand clusters in relation to the original data space. Second, these techniques are not effective in identifying clusters that may exist in different subspaces of the original data space.

Figure 1 gives a 2-dimensional analogue that shows why in general a $K$-means algorithm or its variants may fail to find meaningful clusters in high dimensionality spaces. These algorithms use distance metrics that take all dimensions into account. When the data include sets of dimensions with uniform distributions or noise, this results in clusters that include points that are not close to each other, or clusters that are too small. In the figure, the points are uniformly distributed in the $x$ dimension, but are well clustered on the $y$ dimension. As a result, they form elongated clusters (2 clusters in the example). If we choose $K$ to be to

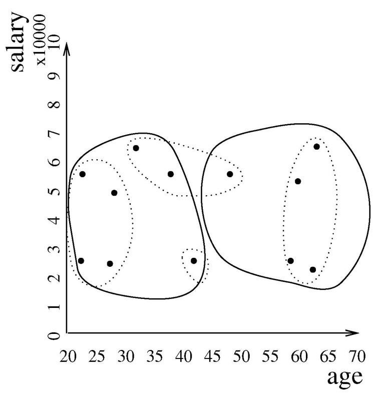

*Figure 1.* $K$-means minimizes the within-cluster variation: $\sum_{i=1}^{K}\sum_{j=1}^{m_i} dist^2(x_j^{(i)}, mean^{(i)})$ where $x_j^{(i)}$ is the $j$-th record in the $i$-th cluster. In this case, because of the uniform distribution in dimension age the clusters found include points that are far apart (clustering for $K = 2$ and 4 is shown).

small (2 in the example) a $K$-means clustering algorithm combines points from different clusters into each of the clusters it produces ; if $K$ is too large (4 in the example), a K-means clustering algorithm partitions each of the data clusters into many small ones. We further discuss these points in the Appendix and Section 4.

Clustering algorithms developed in the database community like BIRCH, CLARANS, and DBSCAN are designed to be scalable, an emphasis not present in the earlier work in the statistics and machine learning literature (Ng and Han, 1994; Zhang et al., 1996). However, these techniques were developed to discover clusters in the full dimensional space. It is not surprising therefore that they are not effective in identifying clusters that exist in subspaces of the original data space. In Section 4, we provide experimental results with BIRCH and DBSCAN in support of this observation.

### *1.2. Contributions and layout of the paper*

We present an algorithm, henceforth referred to as CLIQUE,[1] that satisfies the above desiderata. CLIQUE automatically finds subspaces with high-density clusters. It produces identical results irrespective of the order in which the input records are presented, and it does not presume any canonical distribution for the input data. It generates cluster descriptions in the form of DNF expressions and strives to generate minimal descriptions for ease of comprehension. Empirical evaluation shows that CLIQUE scales linearly with the number of input records, and has good scalability as the number of dimensions (attributes) in the data or the highest dimension in which clusters are embedded is increased.

We begin by formally defining the problem of subspace clustering in Section 2. Section 3, is the heart of the paper where we present CLIQUE. In Section 4, we present a thorough performance evaluation of CLIQUE and conclude with a summary in Section 5.

## 2. Subspace clustering

Before giving a formal description of the problem of subspace clustering, we first give an intuitive explanation of our clustering model.

We are interested in automatically identifying (in general several) subspaces of a high dimensional data space that allow "better" clustering of the data points than the original space. Restricting our search to only subspaces of the original space, instead of using new dimensions (for example linear combinations of the original dimensions), is important because this restriction allows a much simpler, comprehensible presentation of the results. Each of the original dimensions typically has a real meaning to the user, while even a simple linear combination of many dimensions may be hard to interpret (Fayyad et al., 1996).

We use a density based approach to clustering: a cluster is a region that has a higher density of points than its surrounding region. The problem is to automatically identify projections of the input data into a subset of the attributes with the property that these projections include regions of high density.

To approximate the density of the data points, we partition the data space and find the number of points that lie inside each cell (unit) of the partitioning. This is accomplished by partitioning each dimension into the same number of equal length intervals. This means that each unit has the same volume, and therefore the number of points inside it can be used to approximate the density of the unit.

Once the appropriate subspaces are found, the task is to find clusters in the corresponding projections. The data points are separated according to the valleys of the density function. The clusters are unions of connected high density units within a subspace. To simplify their descriptions, we constrain the clusters to be axis-parallel hyper-rectangles.

Each unit in a $k$-dimensional subspace can be described as a conjunction of inequalities because it is the intersection of $2k$ axis-parallel halfspaces defined by $k$ 1-dimensional intervals. Since each cluster is a union of such cells, it can be described with a DNF expression. A compact description is obtained by covering a cluster with a minimal number of maximal, possibly overlapping rectangles and describing the cluster as a union of these rectangles.

Our notion of subspace clustering is tolerant of missing values in input data. A data point is considered to belong to a particular subspace if the attribute values in this subspace are not missing, irrespective of the values of the rest of the attributes. This allows records with missing values to be used for clustering with more accurate results than replacing missing values with values taken from a distribution.

### 2.1. *Problem statement*

Let $\mathcal{A} = \{A_1, A_2, \ldots, A_d\}$ be a set of bounded, totally ordered domains and $\mathcal{S} = A_1 \times A_2 \times \cdots \times A_d$ a $d$-dimensional numerical space. We will refer to $A_1, \ldots, A_d$ as the *dimensions* (attributes) of $\mathcal{S}$.

The input consists of a set of $d$-dimensional points $V = \{v_1, v_2, \ldots, v_m\}$ where $v_i = \langle v_{i1}, v_{i2}, \ldots, v_{id} \rangle$. The $j$th component of $v_i$ is drawn from domain $A_j$.

We partition the data space $\mathcal{S}$ into non-overlapping rectangular *units*. The units are obtained by partitioning every dimension into $\xi$ intervals of equal length; $\xi$ is an input parameter.

Each unit $u$ is the intersection of one interval from each attribute. A unit $u$ has the form $\{u_1, \ldots, u_d\}$ where $u_i = [l_i, h_i)$ is a right-open interval in the partitioning of $A_i$.

We say that a point $v = \langle v_1, \ldots, v_d \rangle$ is contained in a unit $u = \{u_1, \ldots, u_d\}$ if $l_i \leq v_i < h_i$ for all $u_i$. The *selectivity* of a unit is defined to be the fraction of total data points contained in the unit. We call a unit $u$ *dense* if selectivity($u$) is greater than $\tau$, where the *density threshold* $\tau$ is another input parameter.

We similarly define units in all subspaces of the original $d$-dimensional space. Consider a projection of the data set $V$ into $A_{t_1} \times A_{t_2} \times \cdots \times A_{t_k}$, where $k < d$ and $t_i < t_j$ if $i < j$. A unit in the subspace is the intersection of an interval from each of the $k$ attributes.

A *cluster* is a maximal set of connected dense units in $k$-dimensions. Two $k$-dimensional units $u_1, u_2$ are *connected* if they have a common face or if there exists another $k$-dimensional unit $u_3$ such that $u_1$ is connected to $u_3$ and $u_2$ is connected to $u_3$. Units $u_1 = \{r_{t_1}, \ldots, r_{t_k}\}$ and $u_2 = \{r'_{t_1}, \ldots, r'_{t_k}\}$ have a common face if there are $k - 1$ dimensions, assume dimensions $A_{t_1}, \ldots, A_{t_{k-1}}$, such that $r_{t_j} = r'_{t_j}$ and either $h_{t_k} = l'_{t_k}$ or $h'_{t_k} = l_{t_k}$, for $j \in \{1, \ldots, k-1\}$.

A *region* in $k$ dimensions is an axis-parallel rectangular $k$-dimensional set. We are only interested in those regions that can be expressed as unions of units; henceforth all references

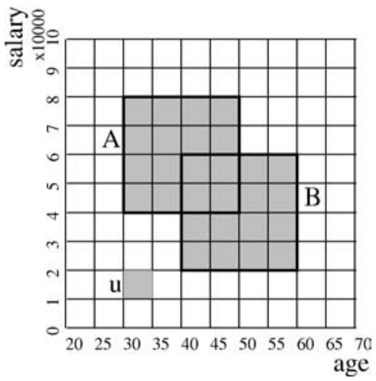

*Figure 2.* Illustration of definitions.

to a region mean such unions. A region can be expressed as a DNF expression on intervals of the domains $A_i$.

We say that a region $R$ is *contained* in a cluster $C$ if $R \cap C = R$. A region $R$ contained in a cluster $C$ is said to be *maximal* if no proper superset of $R$ is contained in $C$.

A *minimal description* of a cluster is a non-redundant covering of the cluster with maximal regions. That is, a minimal description of a cluster $C$ is a set $\mathcal{R}$ of maximal regions such that their union equals $C$ but the union of any proper subset of $\mathcal{R}$ does not equal $C$.

We can now formally define the problem of subspace clustering:

***Subspace clustering.*** Given a set of data points and the input parameters, $\xi$ and $\tau$, find clusters in all subspaces of the original data space and present a minimal description of each cluster in the form of a DNF expression.

***Examples.*** In figure 2, the two dimensional space (age, salary) has been partitioned by a $10 \times 10$ grid. A unit is the intersection of intervals; an example is the unit $u = (30 \leq \text{age} < 35) \wedge (1 \leq \text{salary} < 2)$. A region is a rectangular union of units. $A$ and $B$ are both regions: $A = (30 \leq \text{age} < 50) \wedge (4 \leq \text{salary} < 8)$ and $B = (40 \leq \text{age} < 60) \wedge (2 \leq \text{salary} < 6)$. Assuming that the dense units have been shaded, $A \cup B$ is a cluster. Note that $A$ is a maximal region contained in this cluster, whereas $A \cap B$ is not a maximal region. The minimal description for this cluster is the DNF expression: $((30 \leq \text{age} < 50) \wedge (4 \leq \text{salary} < 8)) \vee ((40 \leq \text{age} < 60) \wedge (2 \leq \text{salary} < 6))$.

In figure 3, assuming $\tau = 20\%$, no 2-dimensional unit is dense and there are no clusters in the original data space. If the points are projected on the salary dimension however, there are three 1-dimensional dense units. Two of these are connected, so there are two clusters in the 1-dimensional salary subspace: $C' = (5 \leq \text{salary} < 7)$ and $D' = (2 \leq \text{salary} < 3)$. There is no cluster in the age subspace because there is no dense unit in that subspace.

***Remarks.*** Our model can be generalized to allow different values of $\xi$ for different dimensions. This will require $\tau$ to be scaled to account for the difference in relative volumes when checking for the density of units in different subspaces.

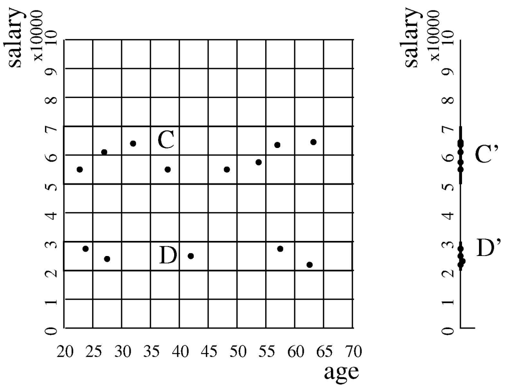

*Figure 3.* Identification of clusters in subspaces (projections) of the original data space.

Our model can also be adapted to handle categorical data. An arbitrary order is introduced in the categorical domain. The partitioning scheme admits one categorical value in each interval and also places an empty interval between two different values. Consequently, if this dimension is chosen for clustering, the clusters will have the same value in this dimension.

Finally, our model for clustering points in a given subspace can be considered nonparametric. The parameters we use, the number of intervals per dimension $\xi$ and the density threshold $\tau$, are only used to approximate the density of the space. We do not presume specific mathematical forms for data distribution; instead, data points are separated according to the valleys of the density function.

***Related work.*** A similar approach to clustering high dimensional data has been proposed by Shoshani (1997). The technique computes an approximation of the density function in a *user-specified* subspace using a grid and then uses this function to cluster the data. On the other hand, we automatically discover the interesting subspaces, and also generate minimal descriptions for the clusters.

A different technique to find rectangular clusters of high density in a projection of the data space has been proposed by Friedman (1997). This algorithm works in a top down fashion. Starting from the full space, it greedily chooses which projection should be taken and reevaluates the solution after each step in order to get closer to an optimal solution.

The subspace identification problem is related to the problem of finding quantitative association rules that also identify interesting regions of various attributes (Srikant and Agrawal, 1996; Miller and Yang, 1997). However, the techniques proposed are quite different. One can also imagine adapting a tree-classifier designed for data mining (e.g. (Mehta et al., 1996; Shafer et al., 1996)) for subspace clustering. In the tree-growth phase, the splitting criterion will have to be changed so that some clustering criterion (e.g., average cluster diameter) is optimized. In the tree-pruning phase, we now minimize the total description length of the clusters obtained and the data description using these clusters.

Recently, Aggarwal and Yu (1999, 2000) have proposed a different technique for identifying clusters in subspaces, PROCLUS. Their technique is similar to iterative clustering techniques (such as the $K$-means or $K$-medoids algorithms described above). The technique iteratively groups the objects into clusters, and eliminates the least relevant dimensions from each of the clusters. Thus their clustering model is not density based as is the one proposed here, and the clustering results can differ. We note that since PROCLUS optimises a criterion similar to the $K$-means algorithm, it can find only more-or-less spherically shaped clusters. Both the number of clusters and the average number of dimensions per cluster are user-defined parameters. ORCLUS (Aggarwal and Yu, 2000) modifies the PROCLUS algorithm by adding a merging process of clusters, and selecting for each cluster principal components instead of attributes.

LAC (Domeniconi et al., 2004) is a recently proposed algorithm that discovers clusters in subspaces spanned by different combinations of dimensions via local weightings of features. This approach mitigates the risk of loss of information encountered in feature selection or global dimensionality reduction techniques.

Recently (Procopiuc et al., 2002), another density-based projective clustering algorithm (DOC/FastDOC) has been proposed. This approach requires the maximum distance between attribute values (i.e. maximum width of the bounding hypercubes) as parameter in input, and pursues an optimality criterion defined in terms of density of each cluster in its corresponding subspace. A Monte Carlo procedure is then developed to approximate with high probability an optimal projective cluster. In practice it may be difficult to set the parameters of DOC, as each relevant attribute can have a different local variance.

## 3. Algorithms

Our clustering technique, CLIQUE, consists of the following steps:

1. Identification of subspaces that contain clusters.
2. Identification of clusters.
3. Generation of minimal description for the clusters.

We discuss algorithms for each of these steps in this section.

### *3.1. Identification of subspaces that contain clusters*

The difficulty in identifying subspaces that contain clusters lies in finding dense units in different subspaces.

***3.1.1. A bottom-up algorithm for finding dense units.*** The simplest way to identify dense units would be to create a histogram in all subspaces and count the points contained in each unit. This approach is infeasible for high dimensional data. We use a bottom-up algorithm that exploits the monotonicity of the clustering criterion with respect to dimensionality to prune the search space. This algorithm is similar to the Apriori algorithm for mining Association rules (Aggarwal et al., 1996). A somewhat similar bottom-up scheme was also used in Chhikara and Register (1979) for determining modes in high dimensional histograms.

**Lemma 1** (Monotonicity). *If a collection of points $S$ is a cluster in a $k$-dimensional space, then $S$ is also part of a cluster in any $(k-1)$-dimensional projections of this space.*

**Proof:** A $k$-dimensional cluster $C$ includes the points that fall inside a union of $k$-dimensional dense units. Since the units are dense, the selectivity of each one is at least $\tau$. All the projections of any unit $u$ in $C$ have at least as large selectivity, because they include all points inside $u$, and therefore are also dense. Since the units of the cluster are connected, their projections are also connected. It follows that the projections of the points in $C$ lie in the same cluster in any $(k-1)$-dimensional projection. □

**Algorithm** The algorithm proceeds level-by-level. It first determines 1-dimensional dense units by making a pass over the data. Having determined $(k-1)$-dimensional dense units, the candidate $k$-dimensional units are determined using the candidate generation procedure given below. A pass over the data is made to find those candidate units that are dense. The algorithm terminates when no more candidates are generated.

The candidate generation procedure takes as an argument $D_{k-1}$, the set of all $(k-1)$-dimensional dense units. It returns a superset of the set of all $k$-dimensional dense units. Assume that the relation $<$ represents lexicographic ordering on attributes. First, we self-join $D_{k-1}$, the join condition being that units share the first $k-2$ dimensions. In the pseudo-code given below for this join operation, $u.a_i$ represents the $i$th dimension of unit $u$ and $u.[l_i, h_i)$ represents its interval in the $i$th dimension.

```
insert into $C_k$
select $u_1.[l_1, h_1), u_1.[l_2, h_2), \ldots,$
    $u_1.[l_{k-1}, h_{k-1}), u_2.[l_{k-1}, h_{k-1})$
from $D_{k-1}\ u_1, D_{k-1}\ u_2$
where $u_1.a_1 = u_2.a_1, u_1.l_1 = u_2.l_1, u_1.h_1 = u_2.h_1,$
    $u_1.a_2 = u_2.a_2, u_1.l_2 = u_2.l_2, u_1.h_2 = u_2.h_2, \ldots,$
    $u_1.a_{k-2} = u_2.a_{k-2}, u_1.l_{k-2} = u_2.l_{k-2}, u_1.h_{k-2} = u_2.h_{k-2},$
    $u_1.a_{k-1} < u_2.a_{k-1}$
```

We then discard those dense units from $C_k$ which have a projection in $(k-1)$-dimensions that is not included in $C_{k-1}$ (Figure 4). The correctness of this procedure follows from the property that for any $k$-dimensional dense unit, its projections in any of $k-1$ dimensions must also be dense (Lemma 1).

*Scalability.* The only phase of CLIQUE in which database records are accessed is the dense unit generation. During the generation of $C_k$, we need storage for dense units $D_{k-1}$ and the candidate units $C_k$. While making a pass over the data, we need storage for $C_k$ and at least one page to buffer the database records. Thus, the algorithm can work with databases of any size. However, memory needs to be managed carefully as the candidates

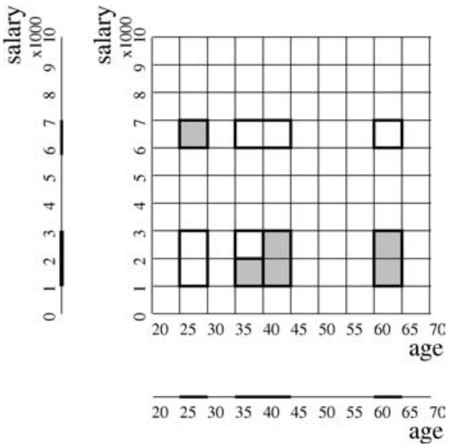

*Figure 4.* The candidate 2-dimensional dense units are computed from the 1-dimensional dense units. The selectivity of each candidate is computed in one database pass (the dense ones are shown filled in the figure).

may swamp the available buffer. This situation is handled by employing a scheme used by Aggarwal et al. (1996). As many candidate units of $C_k$ are generated as will fit in the buffer, and the database is scanned to determine the selectivity of these candidates. Dense units resulting from these candidates are written to disk, while non-dense candidates are deleted. This procedure is repeated until all of $C_k$ has been examined.

*Time complexity.* If a dense unit exists in $k$ dimensions, then all of its projections in a subset of the $k$ dimensions that is, $O(2^k)$ different combinations, are also dense. The running time of our algorithm is therefore exponential in the highest dimensionality of any dense unit. As in Aggarwal et al. (1996), and Gunopulos et al. (1997), it can be shown that the candidate generation procedure produces the minimal number of candidates that can guarantee that all dense units will be found.

Let $k$ be the highest dimensionality of any dense unit and $m$ the number of the input points. The algorithm makes $k$ passes over the database. It follows that the running time of our algorithm is $O(c^k + m\ k)$ for a constant $c$.

The number of database passes can be reduced by adapting ideas from Toivonen (1996) or Brin et al. (1997).

***3.1.2. Making the bottom-up algorithm faster.*** While the procedure just described dramatically reduces the number of units that are tested for being dense, we still may have a computationally infeasible task at hand for high dimensional data. As the dimensionality of the subspaces considered increases, there is an explosion in the number of dense units, and so we need to prune the pool of candidates. The pruned set of dense units is then used to form the candidate units in the next level of the dense unit generation algorithm. The objective is to use only the dense units that lie in "interesting" subspaces.

*MDL-based pruning.* To decide which subspaces (and the corresponding dense units) are interesting, we apply the MDL (Minimal Description Length) principle. The basic idea underlying the MDL principle is to encode the input data under a given model and select the encoding that minimizes the code length (Rissanen, 1989).

Assume we have the subspaces $S_1, S_2, \ldots, S_n$. Our pruning technique first groups together the dense units that lie in the same subspace. Then, for each subspace, it computes the fraction of the database that is covered by the dense units in it: $x_{S_j} = \sum_{u_i \in S_j} count(u_i)$ where $count(u_i)$ is the number of points that fall inside $u_i$. The number $x_{S_j}$ will be referred to as the *coverage* of subspace $S_j$.

Subspaces with large coverage are selected and the rest are pruned. The rationale is that if a cluster exists in $k$ dimensions, then for every subspace of these $k$ dimensions there exist dense units in this subspace (the projections of the dense units that cover the cluster in the original $k$ dimensions) that cover at least the points in the cluster.

We sort the subspaces in the descending order of their coverage. We want to divide the sorted list of subspaces into two sets: the selected set $I$ and the pruned set $P$. The following model is used to arrive at the cut point (see figure 5 for an illustration). For each set, we compute the mean of the cover fractions, and for each subspace in that set we compute the difference from the mean. The code length is the sum of the bit lengths of the numbers we have to store. If we decide to prune subspaces $S_{i+1}, \ldots S_n$, the two averages are $\mu_I(i) = \lceil (\sum_{1 \leq j \leq i} x_{S_j})/i \rceil$ and $\mu_P(i) = \lceil (\sum_{i+1 \leq j \leq n} x_{S_j})/(n-i) \rceil$. Since both $\mu_I(i)$ and $\mu_P(i)$ are integers, the number of bits required to store them is $\log_2(\mu_I(i))$, and $\log_2(\mu_P(i))$ respectively. For each subspace we have to store its difference from $\mu_I(i)$ or $\mu_P(i)$, which is another integer. The total length of the encoding is:

$$\begin{aligned} CL(i) = {} & \log_2(\mu_I(i)) + \sum_{1 \leq j \leq i} \log_2(|x_{S_j} - \mu_I(i)|) \\ & + \log_2(\mu_P(i)) + \sum_{i+1 \leq j \leq n} \log_2(|x_{S_j} - \mu_P(i)|) \end{aligned}$$

This code length is minimized to determine the optimal cut point $i$.

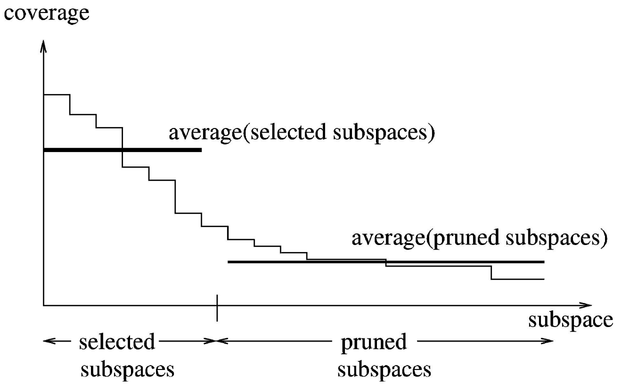

*Figure 5.* Partitioning of the subspaces into selected and prune sets.

*Time complexity.* The optimal cut will be one of the $n - 1$ positions along the sorted sequence, so there are only $n - 1$ sets of pruned subspaces to consider. After sorting, the optimal cut can be computed in two passes of the sorted sequence: In the first pass, we compute $\mu_I(i)$, $\mu_P(i)$ for $1 < i < n$. These averages are used in the second pass to compute $CL(i)$ for $1 < i < n$.

*Remark.* The pruning of dense units in the subspaces with low coverage makes our algorithm faster, but there is a tradeoff because we may now miss some clusters. If a cluster exists in $k$ dimensions, then all of its projections in a subset of the $k$ dimensions are also clusters. In our bottom-up approach, all of them have to considered if we want to find the cluster in $k$ dimensions, but some of them may be in one of the pruned subspaces.

### *3.2. Finding clusters*

The input to the next step of CLIQUE is a set of dense units $D$, all in the same $k$-dimensional subspace $S$. The output will be a partition of $D$ into $D^1, \ldots, D^q$, such that all units in $D^i$ are connected and no two units $u^i \in D^i$, $u^j \in D^j$ with $i \neq j$ are connected. Each such partition is a cluster according to our definition.

The problem is equivalent to finding connected components in a graph defined as follows: Graph vertices correspond to dense units, and there is an edge between two vertices if and only if the corresponding dense units have a common face.

Units corresponding to vertices in the same connected component of the graph are connected because there is a path of units that have a common face between them, therefore they are in the same cluster. On the other hand, units corresponding to vertices in different components cannot be connected, and therefore cannot be in the same cluster.

We use a depth-first search algorithm (Aho et al., 1974) to find the connected components of the graph. We start with some unit $u$ in $D$, assign it the first cluster number, and find all the units it is connected to. Then, if there still are units in $D$ that have not yet been visited, we find one and repeat the procedure. The algorithm is given below:

```
input:    starting unit $u = \{[l_1, h_1), \ldots, [l_k, h_k)\}$
          clusternumber $n$

dfs($u, n$)
$u.num = n$
for ($j = 1; j < k; j{+}{+}$) do begin
    // examine the left neighbor of $u$ in dimension $a_j$
    $u^l = \{[l_1, h_1), \ldots, [(l^l_j), (h^l_j)), \ldots, [l_k, h_k)\}$
    if ($u^l$ is dense) and ($u^l.num$ is undefined)
        dfs($u^l, n$)
    // examine the right neighbor of $u$ in dimension $a_j$
    $u^r = \{[l_1, h_1), \ldots, [(l^r_j), (h^r_j)), \ldots, [l_k, h_k)\}$
    if ($u^r$ is dense) and ($u^r.num$ is undefined)
        dfs($u^r, n$)
end
```

*Time complexity.* The number of dense units for a given subspace cannot be very large, because each dense unit must have selectivity at least $\tau$. We assume therefore that the dense units in this and subsequent steps of CLIQUE can be stored in memory.

We give asymptotic running times in terms of dense unit accesses; the dense units are stored in a main memory data structure (hash tree (Aggarwal et al., 1996)) that allows efficient querying.

For each dense unit visited, the algorithm checks its $2k$ neighbors to find connected units. If the total number of dense units in the subspace is $n$, the total number of data structure accesses is $2kn$.

### *3.3. Generating minimal cluster descriptions*

The input to this step consists of disjoint sets of connected $k$-dimensional units in the same subspace. Each such set is a cluster and the goal is to generate a concise description for it. To generate a minimal description of each cluster, we would want to cover all the units comprising the cluster with the minimum number of regions such that all regions contain only connected units. For a cluster $C$ in a $k$-dimensional subspace $S$, a set $\mathcal{R}$ of regions in the same subspace $S$ is a *cover* of $C$ if every region $R \in \mathcal{R}$ is contained in $C$, and each unit in $C$ is contained in at least one of the regions in $\mathcal{R}$. The *optimal cover* is the cover with the minimal number of rectangles.

Computing the optimal cover is known to be NP-hard, even in the 2-dimensional case (Masek, 1978; Reckhow and Culberson, 1987). The best approximate algorithm known for the special case of finding a cover of a 2-dimensional rectilinear polygon with no holes produces a cover of size bounded by a factor of 2 times the optimal (Franzblau, 1989). Since this algorithm only works for the 2-dimensional case, it cannot be used in our setting. For the general set cover problem, the best known algorithm for approximating the smallest set cover gives an approximation factor of $\ln n$ where $n$ is the size of the universe being covered (Feige, 1996; Lund and Yannakakis, 1993).

This problem is similar to the problem of constructive solid geometry formulae in solid-modeling (Zhang and Bowyer, 1986). It is also related to the problem of covering marked boxes in a grid with rectangles in logic minimization (e.g. (Hong, 1987)). Some clustering algorithms in image analysis (e.g. (Berger and Regoutsos, 1991; Schroeter and Bigun, 1995; Wharton, 1983)) also find rectangular dense regions. In these domains, datasets are in low dimensional spaces, and the techniques used are computationally too expensive for large datasets of high dimensionality.

Our solution to the problem consists of two steps. We first greedily cover the cluster by a number of maximal rectangles (regions), and then discard the redundant rectangles to generate a minimal cover. We only have to consider maximal regions for the cover of a cluster; for any cover $R$ with $c$ regions, we can find another cover $R'$ with $c$ maximal regions, simply by extending each of the non-maximal regions in $C$.

***3.3.1. Covering with maximal regions.*** The input to this step is a set $C$ of connected dense units in the same $k$-dimensional space $S$. The output will be a set $\mathcal{R}$ of maximal regions such that $\mathcal{R}$ is a cover of $C$. We present a greedy growth algorithm for this task.

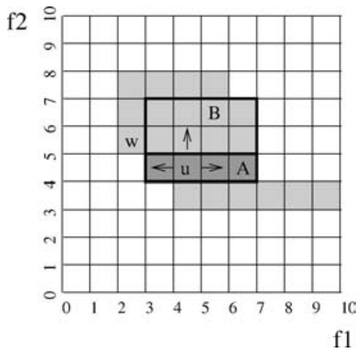

*Figure 6.* Illustration of the greedy growth algorithm.

*Greedy growth.* We begin with an arbitrary dense unit $u_1 \in C$ and greedily grow (as described below) a maximal region $R_1$ that covers $u_1$. We add $R_1$ to $\mathcal{R}$. Then we find another unit $u_2 \in C$ that is not yet covered by any of the maximal regions in $\mathcal{R}$. We greedily grow a maximal region $R_2$ that covers $u_2$. We repeat this procedure until all units in $C$ are covered by some maximal region in $\mathcal{R}$.

To obtain a maximal region covering a dense unit $u$, we start with $u$ and grow it along dimension $a_1$, both to the left and to the right of the unit. We grow $u$ as much as possible in both directions, using connected dense units contained in $C$. The result is a rectangular region. We now grow this region along dimension $a_2$, both to the left and to the right of the region. We again use only connected dense units from $C$, obtaining a possibly bigger rectangular region. This procedure is repeated for all the dimensions, yielding a maximal region covering $u$. The order in which dimensions are considered for growing a dense unit is randomly determined.

Figure 6 illustrates how the algorithm works. Here the dense units appear shaded. Starting from the dense unit $u$, first we grow along the horizontal dimension, finding rectangle $A$ consisting of four dense units. Then $A$ is extended in the vertical dimension. When it cannot be extended further, a maximal rectangle is obtained, in this case $B$. The next step is to find another maximal region starting from a dense unit not covered by $B$, for example $w$.

*Time complexity.* First we show that for each maximal region $R$, the greedy growth algorithm must perform $O(|R|)$ dense unit accesses, where $|R|$ is the number of dense units contained in $R$.

Let $S$ be the subspace that $R$ lies in, $k$ the number of dimensions of $S$, and $n$ the number of dense units in $S$. The greedy growth algorithm must access each unit that a region $R$ covers to ascertain that $R$ is indeed part of a cluster. In addition, it must access every neighbor unit of $R$ to ascertain that $R$ is also maximal. The number of neighbor units is bounded by $2k|R|$, where $|R|$ is the number of dense units contained in $R$.

Since every new maximal region covers at least one thus far uncovered dense unit, the greedy growth algorithm will find at most $O(n)$ new regions. Every new region requires $O(|R|) = O(n)$ dense unit accesses, so the greedy growth algorithm performs a total of $O(n^2)$ dense unit accesses.

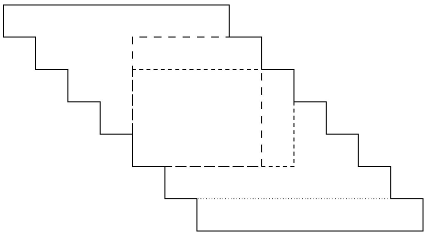

*Figure 7.* The worst case for the greedy growth algorithm (2-dimensional case): Assume $n$ dense units and $n^{1/2}$ upper corners. A minimal cover must include at least one rectangle per upper corner. Since each rectangle is maximal, it must reach the lower staircase as well. This means that the circumference of the rectangle is $2n^{1/2} + 2$, and therefore its area is at least $n^{1/2}$. The sum of the sizes of the rectangles is then $O(n^{2(2-1)/2})$.

This bound is almost tight. Let $S$ contain only one cluster (with $n$ dense units), which is bounded by two parallel hyperplanes and a cylinder which is parallel to only one dimension. Since the hyperplanes are not parallel to any of the $k$ dimensions, the boundary of the cluster that touches the hyperplanes consists of $O(n^{(k-1)/k})$ convex vertices, each of which must be covered by a maximal region. The size of each region is also $O(n^{(k-1)/k})$ since each region has to reach the other hyperplane. In this case the greedy growth algorithm must perform $O(n^{2(k-1)/k})$ dense unit accesses. Figure 7 shows the 2-dimensional analog for this case.

Similarly we show that there can be up to $O(n^{2(k-1)/k})$ maximal regions: we can pair every corner on one hyperplane with every corner on the other, and produce a new maximal region for each pair. The greedy growth algorithm will find $O(n)$ of these.

***3.3.2. Minimal cover.*** The last step of CLIQUE takes as input a cover for each cluster and finds a minimal cover. Minimality is defined in terms of the number of maximal regions (rectangles) required to cover the cluster.

Computing the optimal cover in this case is known to be NP-hard, even in the 2-dimensional case (Masek, 1978; Reckhow and Culberson, 1987). The best approximation algorithm known, gives a factor of 2 for the special case of finding a cover of a 2-dimensional rectilinear polygon with no holes (Franzblau, 1989). If however the polygon has holes, this algorithm only approximates the cover by a factor of $\log\theta$ (where $\theta$ is the size of the optimal cover (Franzblau, 1989)). A similar bound can be derived for the approximation algorithm given by Bronniman and Goodrich (1994). This algorithm assumes that the space we are covering has finite VC-dimension, as is the case here. Note that asymptotically this bound can be as bad as the bound given for the general set cover problem. Computational geometry literature also contains algorithms for covering points with minimum number of rectangles (Franzblau and Kleitman, 1984; Reckhow and Culberson, 1987; Soltan and Gorpinevich, 1992), but they only work for two or three-dimensional datasets. We propose the following greedy heuristic:

*Removal heuristic.* Remove from the cover the smallest (in number of units) maximal region which is redundant (i.e., every unit is also contained in some other maximal region). Break ties arbitrarily. Repeat the procedure until no maximal region can be removed.[2]

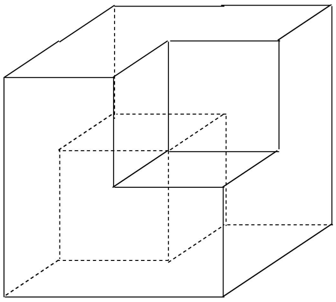

*Figure 8.* An optimal covering consists of 3 regions, each of size $2 \times 1 \times 1$.

*Time complexity.* The removal heuristic is easy to implement and efficient in execution. It needs a simple scan of the sorted list of regions. The cost of sorting the regions is $O(n \log n)$ because the number of dense units $n$ is an upper bound on the number of regions. The scan requires $|R_i|$ dense unit accesses for each region $R_i$. The total number of accesses for all regions is then $\sum |R_i| = O(n^2)$.

Unfortunately, we can prove the following lemma for the worst-case behavior of the removal heuristic:

**Lemma 2.** *In $d$ dimensions, the removal heuristic may construct a cover that is bigger than the minimum by a factor of $d - 1$.*

Consider, for instance, a $2 \times 2 \times 2$ cube, in which 2 opposite corner sub-cubes have been removed. This cluster can be covered by 3 maximal regions (2 solutions), but the removal heuristic may use 4 regions if, for example, the two vertical $1 \times 1 \times 2$ size maximal regions are removed first (see Figure 8.)

The above lower bound suggests that the approximation ratio can, in the worst case, deteriorate with the number of dimensions. However, it is also clear from the lower bound construction that the intricate structure therein is unlikely to arise in practice.

*Stochastic analysis.* We now present a stochastic analysis suggesting that the removal heuristic will do well when each unit is independently dense with probability $p$. This model is quite general: if the number of data points in each unit is a random variable drawn independently from the same (but otherwise arbitrary) distribution as all other units, and we specify any threshold $\tau$ from the domain of these random variables, then each unit is independently dense with some probability $p$ depending on the underlying distribution and $\tau$.

In our application, since a dense unit has high selectivity, $p$ is likely to be small. We show now that provided $p$ is small enough, we obtain a good approximation ratio. Before we prove the theorem, we show the following lemma:

**Lemma 3.** *If $p < 1/8d$, there is a constant $a > 1$ (depending only on $\epsilon = 1/8d - p$) such that the probability that a cluster has size $i$ is at most $a^{-i}$.*

**Proof:** We assume that the dense units in the same subspace are lexicographically ordered. This can be done by ordering the dimensions and, for each dimension, ordering the intervals from least to most significant.

A cluster of $i$ consists of $i$ connected dense units. Since the units are independently dense, the probability of $i$ given units being dense together is $p^i = (1/8d - \epsilon)^i$.

We now have to bound the number of different shapes that $i$ connected units can take.

There are $O(4^i)$ different trees on $i$ vertices (up to isomorphism) (Lovász, 1975). Let us fix such a tree $T$. We can label the nodes of the tree $1, \ldots, i$ so that each $j > 1$ is adjacent to some $i < j$. Given a unit $t$ as a root, we can grow a set of $i$ connected units from $t$ using the tree $T$ and the labeling we created. We traverse $T$ according to the labeling. Each of the units of the connected set corresponds to a node in $T$. The $j$-th unit is added to the connected set when the $j$-th node is encountered. The $j$-th unit must be connected to the unit that corresponds to the parent of the $j$-th node. Since there are at most $2d$ different units that are connected to a given unit, it follows that at each step there are at most $2d$ different ways to grow the tree. Consequently we can create $O((2d)^i)$ different sets of size $i$ that contain a given unit $u$, given one tree with $i$ nodes. There are $O(4^i)$ such trees, so the total number of connected sets, and therefore potential clusters, is $O((8d)^i)$.

Let's assume there is a cluster $c$ with $i$ connected units that includes a unit $u$. Since $c$ is connected, there exists a depth-first-search tree that spans $c$ and is rooted at $u$. By our assumption that there is a lexicographical ordering of the units, this tree is unique.

So our process counts all potential clusters that have size $i$ and contain unit $t$. Since the probability that a given connected set of $i$ units is actually a cluster is $(1/8d - \epsilon)^i$ and since there are $O((8d)^i)$ of them that include a unit $u$, it follows that the expected number of clusters of size $i$ that include $u$ is $(1/(1-8d\epsilon))^{-i}$, and the expected number of clusters of size $i$ is $n(1/(1-8d\epsilon))^{-i}$. The expected number of clusters is then $\sum_i n(1/(1-8d\epsilon))^{-i} = O(n)$.

It follows that the probability of a given cluster having size $i$ is $O(1/(1-8d\epsilon))^{-i}) < \alpha^{-i}$ for some $\alpha > 1$. □

**Theorem 1.** *Let $p = 1/8d - \epsilon$, for any fixed $\epsilon > 0$. If each unit is independently dense with probability at most $p$, then the expected size of the cover we obtain is within a constant factor of the optimal cover.*

**Proof:** Let $c$ be the number of units in a cluster; our algorithm will use at most $c$ maximal rectangles to cover it. To complete the proof, we bound the expected number of maximal rectangles used to cover a given cluster by $\sum_i ia^{-i}$, and the total number of maximal rectangles used to

$$\sum_{\text{clusters}} \sum_i ia^{-i}$$

Let $n$ be the number of clusters found. Since there exists a constant $\theta$, depending only on $a$, such that $\sum_i ia^{-i} \leq \theta$, we have that

$$\sum_{\text{clusters}} \sum_i ia^{-i} \leq \theta\, n$$

It follows that the expected size of the cluster covers we find is within a constant factor of the total number of clusters, and thus the size of the optimal cover. □

## 4. Performance experiments

We now empirically evaluate CLIQUE using synthetic as well as real datasets. The goals of the experiments are to assess the efficiency and accuracy of CLIQUE:

- *Efficiency:* Determine how the running time scales with:
  - Dimensionality of the data space.
  - Dimensionality of clusters.
  - Size of database.
- *Accuracy:* Test if CLIQUE recovers known clusters in some subspaces of a high dimensional data space.

The experiments were run on a 133-MHz IBM RS/6000 Model 43P workstation. The data resided in the AIX file system and was stored on a 2GB SCSI drive with sequential throughput of about 2 MB/second.

### *4.1. Synthetic data generation*

We use the synthetic data generator from Zait and Messatfa (1997) to produce datasets with clusters of high density in specific subspaces. The data generator allows control over the structure and the size of datasets through parameters such as the number of records, the number of attributes, and the range of values for each attribute. We assume a bounded data space ($n$-dimensional cube) that data points live in. The range of values was set to [0, 100] for all attributes. The data space is partitioned into a multidimensional grid generated by dividing each dimension into 100 partitions of equal length. Each box of this grid forms a unit.

The clusters are hyper-rectangles in a subset of dimensions such that the average density of data points inside the hyper-rectangle is much larger than the average density in the subspace. The faces of such a cluster are parallel to the axis, therefore another way to describe the cluster is as the intersection of a set of attribute ranges.

The cluster descriptions are provided by the user. A description specifies the subspace of each hyper-rectangle and the range for each attribute in the subspace. The attribute values for a data point assigned to a cluster are generated as follows. For those attributes that define the subspace in which the cluster is embedded, the value is drawn independently at random from the uniform distribution within the range of the hyper-rectangle. For the remaining attributes, the value is drawn independently at random from the uniform distribution over the entire range of the attribute. After distributing the specified number of points equally among the specified clusters, an additional 10% points are added as random noise. Values for all the attributes of these points are drawn independently at random from the uniform distribution over the entire range of the attribute.

### *4.2. Synthetic data results*

We first present scalability and accuracy results observed using synthetic data. The experiments were run with $\xi = 10$. All times are in seconds.

*Database size.* Figure 9 shows the scalability as the size of the database is increased from 100,000 to 500,000 records. The data space had 50 dimensions and there were 5 clusters, each in a different 5-dimensional subspace, and $\tau$ was set to 0.5%. As expected, the running time scales linearly with the size of the database because the number of passes through the database does not change.

*Dimensionality of the data space.* Figure 10 shows the scalability as the dimensionality of the data space is increased from 10 to 100. The database had 100,000 records and there where 5 clusters, each in a different 5-dimensional subspace, and $\tau$ was set to 0.5%. The curve exhibits quadratic behavior. We note that the problem of searching for interesting subspaces inherently does not scale well as the dimensionality of the data space increases. In this case, we are searching for clusters in 5 dimensions. The number of 5-dimensional

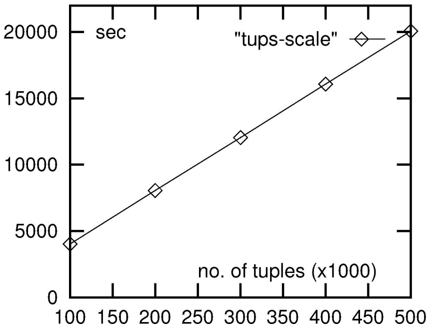

*Figure 9.* Scalability with the number of data records.

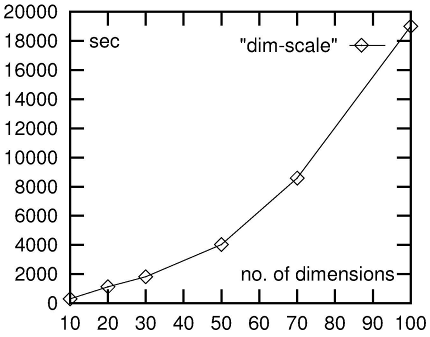

*Figure 10.* Scalability with the dimensionality of the data space.

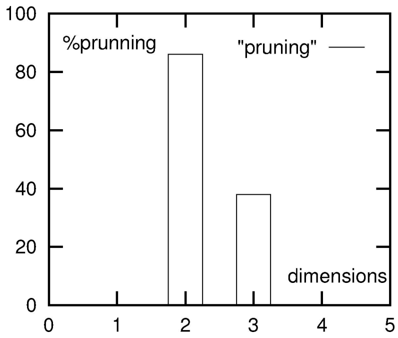

*Figure 11.* Number of subspaces pruned.

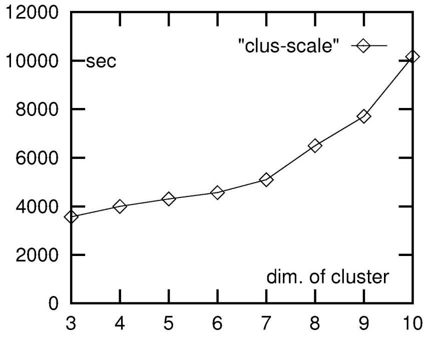

*Figure 12.* Scalability with the dimensionality of the hidden cluster.

subspaces of a $d$-dimensional space is $O(d^5)$. The algorithm performs better than the worst case because many of these dimensions are pruned during the dense unit generation phase.

Figure 11 shows the percentage of subspaces pruned by MDL during an algorithm run. The input was a synthetic dataset with 50 dimensions, with 5 hidden 5-dimensional clusters, and $\tau$ was set to 0.5%. In this case, 86% of the 2-dimensional subspaces and 38% of the 3-dimensional subspaces were pruned. The result of the pruning is a much faster algorithm, though there is now a risk of missing some clusters.

*Dimensionality of hidden clusters.* Figure 12 shows the scalability as the highest dimensionality of the hidden clusters is increased from 3 to 10 in a 50-dimensional space. In each case, one cluster was embedded in the relevant subspace of highest dimensionality. The database had 100,000 records and $\tau$ was set at 1% for 3-dimensional clusters, 0.5% for 4-dimensional to 7-dimensional clusters and and 0.1% for 8-dimensional to 10-dimensional clusters. We selected a lower $\tau$ for the highest dimensional clusters because, as the volume of the clusters increases, the cluster density decreases. For lower dimensions however we can increase $\tau$, and since this does not increase the number of dense units the algorithm runs at least as fast. The increase in running time reflects the time complexity of our algorithm, which is $O(mk + c^k)$ where $m$ is the number of records, $c$ a constant, and $k$ the maximum dimensionality of the hidden clusters.

*Accuracy.* In all the above experiments, the original clusters were recovered by the algorithm. In some cases, a few extra clusters were reported, typically comprising a single dense unit with very low selectivity. This artifact is a byproduct of the data generation algorithm and the fact that $\tau$ was set low. As a result, some units had enough noise points to become dense.

### *4.3. Comparisons with BIRCH, DBSCAN and SVD*

We ran CLIQUE, BIRCH, and DBSCAN with the same synthetic datasets.[3] The purpose of these experiments was to assess if algorithms such as BIRCH or DBSCAN designed for clustering in the full dimensional space can also be used for subspace clustering. For the task of finding clusters in the full dimensional space, which was the design goal of these algorithms, CLIQUE has no advantage.

We used clusters embedded in 5-dimensional subspaces while varying the dimensionality of the space from 5 to 50. For reference, CLIQUE was able to recover all clusters in every case.

*BIRCH.* We provided the correct number of clusters (5) as input to the postprocessing clustering algorithm built on top of BIRCH. The output consists of cluster centers in the full dimensional space. The input datasets had 100,000 points. The input clusters were hyper-rectangles in 5-dimensional subspaces, with the values of the remaining attributes uniformly distributed. This is equivalent to a hyper-rectangle in the full data space where remaining attributes include the whole range. Therefore BIRCH successfully recovers a cluster if it reports a center approximately at the center of the equivalent hyper-rectangle in the full data space, and the number of points in the reported cluster is approximately correct.

The results summarized in Table 1 show that BIRCH can discover 5-dimensional clusters embedded in a 10-dimensional data space, but fails to do so when the dimensionality of the

*Table 1.* BIRCH experimental results.

| Dim. of data | Dim. of clusters | No. of clusters | Clusters found | True clusters identified |
|---|---|---|---|---|
| 5 | 5 | 5 | 5 | 5 |
| 10 | 5 | 5 | 5 | 5 |
| 20 | 5 | 5 | 3, 4, 5 | 0 |
| 30 | 5 | 5 | 3, 4 | 0 |
| 40 | 5 | 5 | 3, 4 | 0 |
| 50 | 5 | 5 | 3 | 0 |

*Table 2.* DBSCAN experimental results.

| Dim. of data | Dim. of clusters | No. of clusters | Clusters found | True clusters identified |
|---|---|---|---|---|
| 5 | 5 | 5 | 5 | 5 |
| 7 | 5 | 5 | 5 | 5 |
| 8 | 5 | 5 | 3 | 1 |
| 10 | 5 | 5 | 1 | 0 |

data space increases. This is expected because BIRCH uses a distance function that takes all dimensions into account. When the number of dimensions with uniform distribution increases, the distance function fails to distinguish the clusters.

As dimensionality of the data space increases, BIRCH does not always return 5 clusters even though 5 is given as an input parameter. For different randomly generated datasets, it returns 3, 4 or 5 clusters. The final column gives the number of correct embedded clusters that BIRCH identified.

*DBSCAN.* DBSCAN discovers the number of clusters on its own, so we did not have to give the number of clusters as input. DBSCAN could not be run with data having more than 10 dimensions. The input datasets had 10,000 points. As in the BIRCH experiments, the clusters were in 5-dimensional subspaces. We ran DBSCAN with different input values of $\epsilon$; we report the best results in Table 2.

DBSCAN could not discover 5-dimensional clusters in a 10-dimensional data space; it could do so when the dimensionality of the space was reduced to 7. Even in a 8-dimensional data space, it could recover only one of the 5-dimensional embedded clusters. DBSCAN uses a density based cluster definition, and even a small number of dimensions with uniform distribution can lower the density in the space enough so that no clusters are found.

*SVD.* We also did Singular Value Decomposition (SVD) (Duda and Hart, 1973; Fukunaga, 1990) on the synthetic datasets to find if the dimensionality of the space can be reduced or if the subspaces that contain dense units can be deduced from the projections into the new space.

In Table 3, $r_k$ gives the ratio of the sum of the $k$ largest eigenvalues to the sum of all eigenvalues. Let $\lambda_1, \ldots, \lambda_d$ be the eigenvalues found, sorted in decreasing order. Then $r_k = \sum_{i=1}^{k} \lambda_i / \sum_{i=1}^{d} \lambda_i$. The quantity $r_k$ indicates how much variance is retained in the new space that is defined by the $k$ eigenvectors corresponding to the $k$ largest eigenvalues. In our experiments the variation of the original space is such that the smallest eigenvalue is almost as large as the largest, and so we cannot achieve any dimensionality reduction. In addition, the new projections are linear combinations of all the original vectors and cannot be used to identify the subspaces that contain clusters.

*Table 3*. SVD decomposition experimental results.

| Dim. of data ($d$) | Dim. of clusters | No. of clusters | $r_{d/2}$ | $r_{(d-5)}$ | $r_{(d-1)}$ |
|---|---|---|---|---|---|
| 10 | 5 | 5 | 0.647 | 0.647 | 0.937 |
| 20 | 5 | 5 | 0.606 | 0.827 | 0.969 |
| 30 | 5 | 5 | 0.563 | 0.858 | 0.972 |
| 40 | 5 | 5 | 0.557 | 0.897 | 0.981 |
| 50 | 5 | 5 | 0.552 | 0.919 | 0.984 |

*Table 4*. Real data experimental results.

| Dataset | Dim. of data | Dim. of clusters | No. of clusters |
|---|---|---|---|
| Insur1 | 9 | 7 | 2 |
| Insur2 | 7 | 4 | 5 |
| Store | 24 | 10 | 4 |
| Bank | 52 | 9 | 1 |

### 4.4. *Real data results*

We ran CLIQUE against two datasets obtained from the insurance industry (Insur1, Insur2), another from a department store (Store), and the last from a bank (Bank). Table 4 summarizes the results of this experiment. Each run used a selectivity threshold of 1%, and every dimension was divided into 10 intervals of equal length. We show in the table the dimensionality of the original data space, the highest dimensionality of the subspace in which clusters were found, and the number of such clusters for each of the datasets. In all cases, we discovered meaningful clusters embedded in lower dimensional subspaces.

## 5. Conclusions

We introduced the problem of automatic subspace clustering, motivated by the needs of emerging data mining applications. The solution we propose, CLIQUE, has been designed to find clusters embedded in subspaces of high dimensional data without requiring the user to guess subspaces that might have interesting clusters. CLIQUE generates cluster descriptions in the form of DNF expressions that are minimized for ease of comprehension. It is insensitive to the order of input records and does not presume some canonical data distribution. In designing CLIQUE, we combined developments from several fields including data mining, stochastic complexity, pattern recognition, and computational geometry.

Empirical evaluation shows that CLIQUE scales linearly with the size of input and has good scalability as the number of dimensions in the data or the highest dimension in which clusters are embedded is increased. CLIQUE was able to accurately discover clusters embedded in lower dimensional subspaces, although there were no clusters in the original data space. Having demonstrated the computational feasibility of automatic subspace clustering, we believe it should be considered a basic data mining operation along with other operations such as associations and sequential-patterns discovery, time-series clustering, and classification (Internationl Business Machines, 1996).

Automatic subspace clustering can be useful in other applications besides data mining. To index OLAP data, for instance, the data space is first partitioned into dense and sparse regions (Earle, 1994). Data in dense regions is stored in an array whereas a tree structure is used to store sparse regions. Currently, users are required to specify dense and sparse dimensions (Arbor Software Corporation, ). Similarly, the precomputation techniques for range queries over OLAP data cubes (Ho et al., 1997) require identification of dense regions in sparse data cubes. CLIQUE can be used for this purpose.

In future work, we plan to address the problem of evaluating the quality of clusterings in different subspaces. One approach is to choose clusters that maximize the ratio of cluster density over expected density for clusterings with the same dimensionality. We also plan to investigate what system support can be provided to the user for selecting the model parameters, $\tau$ and $\xi$. Another area for future work is to try an alternative approach for finding dense units. If the user is only interested in clusters in the subspaces of highest dimensionality, we can use techniques based on recently proposed algorithms for discovering maximal itemsets (Bayardo, 1998; Lin and Kedem, 1998). These techniques will allow CLIQUE to find dense units of high dimensionality without having to find all of their projections.

## Appendix: Dimensionality reduction

The principal component analysis or Karhunen-Loève (KL) transformation is the optimal way to project $n$-dimensional points to $k$-dimensional points such that the error of the projections (the sum of the squared distances) is minimal (Duda and Hart, 1973; Fukunaga, 1990). This transformation gives a new set of orthogonal axes, each a linear combination of the original ones, sorted by the degree by which they preserve the distances of the points in the original space.

For a given set of $m$ points in $d$ dimensions, finding the set of axes in the KL transformation is equivalent to solving the Singular Value Decomposition problem in an $m \times d$ matrix $N$, each row of which represents a data point. The SVD of the matrix $N$ is the decomposition into $N = U \times \Lambda \times V^t$, where $U$ is an $m \times r$ matrix, $\Lambda$ a diagonal $r \times r$ matrix and $V$ a column orthonormal $d \times r$ matrix. The matrix $V$ represents the axes of the KL-decomposition (they are also the eigenvectors of the matrix $N \times N^t$), ordered by the respective values in the matrix $\Lambda$. Note that $r \leq d$, so the new space has potentially lower dimensionality. In addition, for each small entry in the matrix $\Lambda$, the corresponding vectors may be eliminated and a lower dimensionality space obtained.

In our problem, we assume there may not be clearly defined clusters in the original space and try to find those dimensions that can be used for clustering. Clearly, two points may be

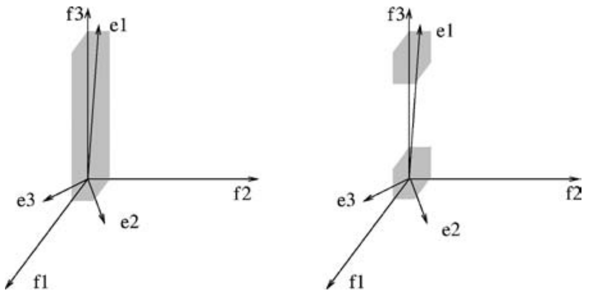

*Figure 13*. Examples where KL transformation is helpful.

far apart in a 3-dimensional space but could be quite close when a specific projection into 2 dimensions is used. The effects of such projections are what we are trying to capture. In addition, in the interest of comprehension, we do not want to use dimensions that are linear combinations of the original ones.

The following examples illustrate these points. In figure 13, two data distributions are shown. The original axes are labeled $f1$, $f2$, $f3$. In both cases, the data points are uniformly distributed inside the shaded area. In the left case there is one cluster, in the right two. Assuming the number of points to be the same in both cases, the density of the shaded regions is different for the two sets. The eigenvectors are labeled $e1$, $e2$, $e3$, such that $e1$ corresponds to the eigenvalue with the largest magnitude and $e3$ to the eigenvalue with the smallest magnitude. The first eigenvalue is much larger than the other two, indicating that there is large variation along axis $e1$. The eigenvectors are essentially the same in both cases. Thus, it can be said that the KL transformation is quite successful in these instances. Although the transformation cannot be used by itself to find the actual clusters because it cannot distinguish between the two cases, one can argue that the clusters will be discovered after projecting the points on $e1$ and examining the distribution of the projections.

In figure 14, the 2-dimensional data is uniformly distributed in dimension $f1$, but contains three clusters along dimension $f2$. Despite the clustering on $f2$, there is large variation

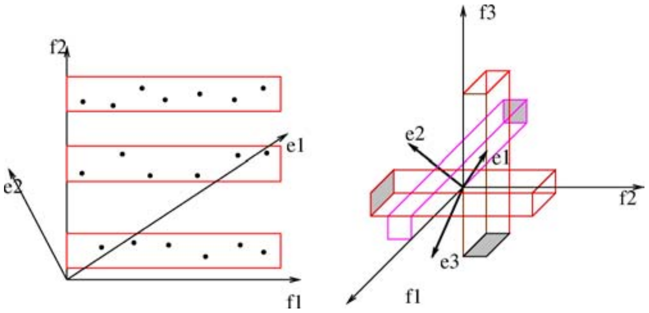

*Figure 14*. Examples where KL transformation is not helpful.

along both axes. The results of the KL transformation are the eigenvectors $e1$ and $e2$ as shown. Because of the variation, the eigenvalue corresponding to eigenvector $e2$ (the second largest eigenvalue) is quite large. We have thus come up with a space of the same dimensionality. Furthermore, no projection on the new axes can be used to identify the clusters.

The right figure illustrates that clusters may exist in different subspaces. The data points are uniformly distributed inside the three 3-dimensional rectangles. The rectangles are long, skinny and not very dense. In addition they do not intersect. For reasonable selectivities, the only clusters are the projections of the rectangles on their small faces; that is, one cluster in each of the $f1 \times f2$, $f1 \times f3$ and $f2 \times f3$ subspaces. The KL decomposition does not help here because of large variation along each of the original axes. The resulting axes are $e1$, $e2$, $e3$ and the three eigenvalues are approximately equal. This means there is no 2-dimensional space which approximates the original space. A 3-dimensional space has to be used after the KL transformation for the clustering. But the density of the points in the 3-dimensional space is too low to obtain good clustering.

## Acknowledgment

The code for CLIQUE builds on several components that Ramakrishnan Srikant wrote for quantitative association rules. Srikant generously explained his code and modified it in places to meet CLIQUE's requirements. Roberto Bayardo provided helpful feedback to improve the presentation of the paper. Christoph Lingenfelder, Stefanos Manganaris, Hammou Messatfa, and Julio Ortega provided the datasets used in this study. We are also thankful to Prof. H.-P. Kriegel, Raymond Ng, and Raghu Ramakrishnan for providing access to DBSCAN, CLARANS, and BIRCH, respectively.

## Notes

1. For CLustering In QUEst, the data mining research project at IBM Almaden. A preliminary version of our results appeared in Aggarwal et al. (1998).
2. We also considered the following *Addition heuristic*: View the cluster as empty space. Add to the cover the maximal region that will cover the maximum number of yet uncovered units in the cluster. Break ties arbitrarily. Repeat the procedure until the whole cluster is covered.

   For general set cover, the addition heuristic is known to give a cover within a factor $\ln n$ of the optimum where $n$ is the number of units to be covered (Lovász, 1975). Thus it would appear that the addition heuristic, since its quality of approximation matches the negative results of Feige (1996), and Lund and Yannakakis (1993), would be the obvious choice. However, its implementation in our high dimensional geometric setting is too inefficient. The implementation requires the rather complex computation of the number of uncovered units a candidate maximal region will cover. The residual uncovered regions that arise as the cover is formed can be complicated, and no efficient data structures are known for efficiently maintaining the uncovered units.
3. We could not run experiments with CLARANS because the code required modification to work with points in high dimensions. We expect CLARANS to show similar behavior as BIRCH in identifying clusters embedded in subspaces.

## References

Aggarwal, C.C. and Yu, P.S. 2000. Finding generalized projected clusters in high dimensional spaces. In Proc. of SIGMOD 2000 Conference, pp. 70–81.

Aggrawal, C., Procopiuc, C., Wolf, J., Yu, P., and Park, J. 1999. Fast algorithms for projected clustering. In Proc. of 1999 ACM SIGMOD Int. Conf. on Management of Data, Philadelphia, PA.

Agrawal, R., Gehrke, J., Gunopulos, D., and Raghavan, P. 1998. Automatic subspace clustering of high dimensional data for data mining applications. In Proc. of 1998 ACM SIGMOD Int. Conf. on Management of Data, pp. 94–105.

Agrawal, R., Mannila, H., Srikant, R., Toivonen, H., and Verkamo, A.I. 1996. Fast discovery of association rules. In Advances in Knowledge Discovery and Data Mining, U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy (Eds.). AAAI/MIT Press, Chap 12, pp. 307–328.

Aho, A., Hopcroft, J., and Ullman, J. 1974. The Design and Analysis of Computer Algorithms. Addison-Welsley.

Arabie, P. and Hubert, L.J. 1996. An overview of combinatorial data analyis. In Clustering and Classification. P. Arabie, L. Hubert, and G.D. Soete, (Eds.). New Jersey: World Scientific Pub., pp. 5–63.

Arbor Software Corporation. Application Manager User's Guide, Essbase Version 4.0 edition.

Bayardo, R. 1998. Efficiently mining long patterns from databases. In Proc. of the ACM SIGMOD Conference on Management of Data, Seattle, Washington.

Berchtold, S., Bohm, C., Keim, D., and Kriegel, H.-P. 1997. A cost model for nearest neighbor search in high-dimensional data space. In Proceedings of the 16th Symposium on Principles of Database Systems (PODS), pp. 78–86.

Berger, M. and Regoutsos, I. 1991. An algorithm for point clustering and grid generation. IEEE Transactions on Systems, Man and Cybernetics, 21(5):1278–86.

Brin, S., Motwani, R., Ullman, J. D., and Tsur, S. 1997. Dynamic itemset counting and implication rules for market basket data. In Proc. of the ACM SIGMOD Conference on Management of Data.

Bronniman, H. and Goodrich, M. 1994. Almost optimal set covers in finite VC-dimension. In Proc. of the 10th ACM Symp. on Computational Geometry, pp. 293–302.

Cheeseman, P. and Stutz, J. 1996. Bayesian classification (autoclass): Theory and results. In Advances in Knowledge Discovery and Data Mining. U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, (Eds.). Chap 6. AAAI/MIT Press, pp. 153–180.

Chhikara, R. and Register, D. 1979. A numerical classification method for partitioning of a large multidimensional mixed data set. Technometrics, 21:531–537.

Domeniconi, C., Papadopoulos, D., Gunopulos, D., and Ma, S. 2004. Subspace clustering of high dimensional data. SIAM International Conference on Data Mining (SDM).

Duda, R.O. and Hart, P.E. 1973. Pattern Classification and Scene Analysis. John Wiley and Sons.

Earle, R.J. 1994. Method and apparatus for storing and retrieving multi-dimensional data in computer memory. U.S. Patent No. 5359724.

Ester, M., Kriegel, H.-P., Sander, J., and Xu, X. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In Proc. of the 2nd Int'l Conference on Knowledge Discovery in Databases and Data Mining, Portland, Oregon.

Ester, M., Kriegel, H. -P., and Xu, X. 1995. A database interface for clustering in large spatial databases. In Proc. of the 1st Int'l Conference on Knowledge Discovery in Databases and Data Mining, Montreal, Canada.

Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P., and Uthurusamy, R. (Eds.). 1996. Advances in Knowledge Discovery and Data Mining. AAAI/MIT Press.

Feige, U. 1996. A threshold of ln n for approximating set cover. In Proceedings of the Twenty-Eighth Annual ACM Symposium on Theory of Computing, pp. 314–318.

Franzblau, D. 1989. Performance guarantees on a sweep-line heuristic for covering rectilinear polygons with rectangles. SIAM J. Disc. Math, 2(3):307–321.

Franzblau, D.S. and Kleitman, D.J. 1984. An algorithm for constructing regions with rectangles: Independence and minimum generating sets for collections of intervals. In Proc. of the 6th Annual Symp. on Theory of Computing, Washington D.C., pp. 268–276.

Friedman, J. 1997. Optimizing a noisy function of many variables with application to data mining. In UW/MSR Summer Research Institute in Data Mining.

Fukunaga, K. 1990. Introduction to Statistical Pattern Recognition. Academic Press.

Guha, S., Rastogi, R., and Shim, K. 1998. CURE: An efficient clustering algorithm for large databases. Proceedings of ACM SIGMOD, pp. 73–84.

Gunopulos, D., Khardon, R., Mannila, H., and Saluja, S. 1997. Data mining, hypergraph transversals, and machine learning. In Proc. of the 16th ACM Symp. on Principles of Database Systems, pp. 209–216.

Ho, C.-T., Agrawal, R., Megiddo, N., and Srikant, R. 1997. Range queries in OLAP data cubes. In Proc. of the ACM SIGMOD Conference on Management of Data, Tucson, Arizona.

Hong, S.J. 1987. MINI: A heuristic algorithm for two-level logic minimization. In Selected Papers on Logic Synthesis for Integrated Circuit Design, R. Newton (Eds.). IEEE Press.

Internationl Business Machines. 1996. IBM Intelligent Miner User's Guide, Version 1 Release 1, SH12-6213-00 edition, July 1996.

Jain, A.K. and Dubes, R.C. 1988. Algorithms for Clustering Data. Prentice Hall.

Kaufman, L. and Rousseeuw, P. 1990. Finding Groups in Data: An Introduction to Cluster Analysis. John Wiley and Sons.

Lin, D.-I. and Kedem, Z.M. 1998. Pincer search: A new algorithm for discovering the maximum frequent sets. In Proc. of the 6th Int'l Conference on Extending Database Technology (EDBT), Valencia, Spain.

Lovász, L. 1975. On the ratio of the optimal integral and fractional covers. Discrete Mathematics, 13:383–390.

Lund, C. and Yannakakis, M. 1993. On the hardness of approximating minimization problems. In Proceedings of the ACM Symposium on Theory of Computing, pp. 286–293.

Masek, W. 1978. Some NP-Complete Set Covering Problems. M.S. Thesis, MIT.

Mehta, M., Agrawal, R., and Rissanen, J. 1996. SLIQ: A fast scalable classifier for data mining. In Proc. of the Fifth Int'l Conference on Extending Database Technology (EDBT), Avignon, France.

Michalski, R.S. and Stepp, R.E. 1983. Learning from observation: Conceptual clustering. In Machine Learning: An Artificial Intelligence Approach, R.S. Michalski, J.G. Carbonell, and T. M. Mitchell (Eds.). Volume I. Morgan Kaufmann, pp. 331–363.

Miller, R. and Yang, Y. 1997. Association rules over interval data. In Proc. ACM SIGMOD International Conf. on Management of Data, pp. 452–461.

Ng, R.T. and Han, J. 1994. Efficient and effective clustering methods for spatial data mining. In Proc. of the VLDB Conference, Santiago, Chile.

Procopiuc, C.M., Jones, M., Agarwal, P.K., and Murali, T.M. 2002. A Monte Carlo algorithm for fast projective clustering. SIGMOD.

Reckhow, R.A. and Culberson, J. 1987. Covering simple orthogonal polygon with a minimum number of orthogonally convex polygons. In Proc. of the ACM 3rd Annual Computational Geometry Conference, pp. 268–277.

Rissanen, J. 1989. Stochastic Complexity in Statistical Inquiry. World Scientific Publ. Co.

Schroeter, P. and Bigun, J. 1995. Hierarchical image segmentation by multi-dimensional clustering and orientation-adaptive boundary refinement. Pattern Recognition, 25(5):695–709.

Shafer, J., Agrawal, R. and Mehta, M. 1996. SPRINT: A scalable parallel classifier for data mining. In Proc. of the 22nd Int'l Conference on Very Large Databases, Bombay, India.

Shoshani, A. Personal communication, 1997.

Sneath, P. and Sokal, R. 1973. Numerical Taxonomy. Freeman.

Soltan, V. and Gorpinevich, A. 1992. Minimum dissection of rectilinear polygon with arbitrary holes into rectangles. In Proc. of the ACM 8th Annual Computational Geometry Conference, Berlin, Germany, pp. 296–302.

Srikant, R. and Agrawal, R. 1996. Mining quantitative association rules in large relational tables. In Proc. of the ACM SIGMOD Conference on Management of Data, Montreal, Canada.

Toivonen, H. 1996. Sampling large databases for association rules. In Proc. of the 22nd Int'l Conference on Very Large Databases, Mumbai (Bombay), India, pp. 134–145.

Wharton, S. 1983. A generalized histogram clustering for multidimensional image data. Pattern Recognition, 16(2):193–199.

Zait, M. and Messatfa, H. 1997. A comparative study of clustering methods. Future Generation Computer Systems, 13(2-3):149–159.

Zhang, D. and Bowyer, A. 1986. CSG set-theoretic solid modelling and NC machining of blend surfaces. In Proceedings of the Second Annual ACM Symposium on Computational Geometry, pp. 314–318.

Zhang, T., Ramakrishnan, R., and Livny, M. 1996. BIRCH: An efficient data clustering method for very large databases. In Proc. of the ACM SIGMOD Conference on Management of Data, Montreal, Canada.