On Exploring Complex Relationships of Correlation Clusters

Elke Achtert, Christian Böhm, Hans-Peter Kriegel,
Peer Kröger, Arthur Zimek

Institute for Informatics
Ludwig-Maximilians-Universität München
Germany

SSDBM 07

Outline

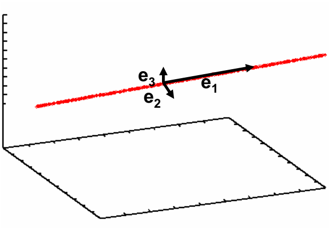
1. Introduction
1.1 Correlation Clusters
1.2 Related Work
1.3 Complex Relationships

2. Algorithm ERiC
2.1 Partitioning w.r.t. Correlation Dimensionality
2.2 Computing Correlation Clusters within each Partition
2.3 Aggregating the Hierarchy of Correlation Clusters

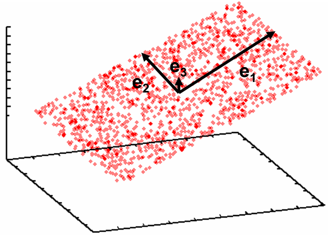
3. Evaluation
3.1 Efficiency
3.2 Effectivity

4. Conclusions

- In high-dimensional data, meaningful clusters are usually based only on a subset of all dimensions.
  - subspace/projected clustering: axis parallel subspaces (2^d possibilities)
  - arbitrarily oriented subspaces (infinite, uncountable possibilities)



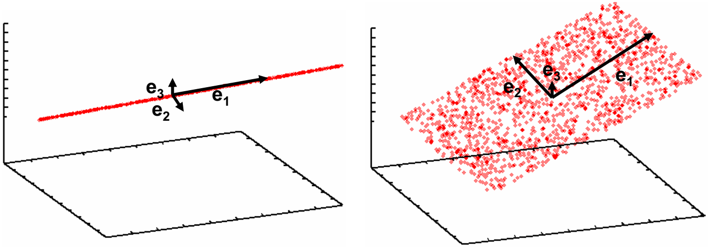
(a) dense cluster in span(v_2,v_3)



(b) dense cluster in span(v_3)

What are Correlation Clusters?

- The term “correlation cluster” highlights the opposite viewpoint on finding arbitrarily oriented subspace clusters:



(a) $\lambda = 1$ (b) $\lambda = 2$

- The subspace orthogonal to the subspace, where the points cluster very dense, appears as a (λ-dimensional) hyperplane accomodating many data points with a high variance.
- This hyperplane indicates complex linear relationships among the attributes contributing to a base of the hyperplane.

Formalizing Correlation Clusters

- derive the local covariance matrix $\Sigma_{\mathcal{C}}$ for cluster $\mathcal{C}$ (or for a representative set, e.g. the local neighborhood of a point)
- decomposition (PCA) of $\Sigma_{\mathcal{C}}$ to eigenvalues E and eigenvectors V
- most of the variance covered by small number of eigenvectors
- number of eigenvectors covering most of the variance is called correlation dimensionality of a cluster $\mathcal{C}$: $\lambda_{\mathcal{C}}$
- eigenvectors $\#1 \ldots \#\lambda_{\mathcal{C}}$: strong eigenvectors
- eigenvectors $\#\lambda_{\mathcal{C}} + 1 \ldots \#d$: weak eigenvectors
- selection matrix for weak eigenvectors: $\hat{E}$ with entries $\hat{e}_{ij} \in \{0, 1\}$, $i, j = 1, \ldots, d$:

$$\hat{e}_{ij} = \begin{cases} 1 & if \quad i = j > \lambda_p \\ 0 & otherwise \end{cases}$$

- weak eigenvectors: $V \cdot \hat{E}$

Other Approaches

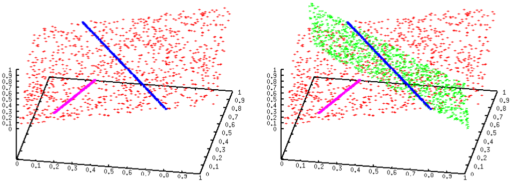
Several approaches to correlation clustering facilitate PCA to derive local similarity measures.

- ORCLUS [Aggarwal, Yu (SIGMOD 2000)] incorporates PCA into a k-means-like approach – drawback: user needs to specify number of clusters in advance
- 4C [Böhm et al. (SIGMOD 2004)] integrates PCA into density-based clustering – drawback: user needs to specify global density threshold

Both tend to find clusters of a dimensionality close to a user specified value, instead of detecting all correlation clusters hidden in the data set.

- HiCO [Achtert et al. (SSDBM 2006)] uses hierarchical clustering to find correlation clusters over a broad range of intrinsic dimensionalities – drawbacks:
  - very expensive procedure
  - limited to strict hierarchies

Complex Relationships

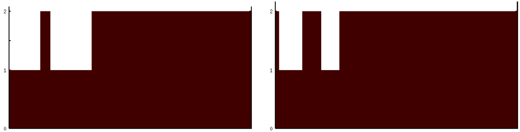


(a) Sample data set 1 (b) Sample data set 2

Figure: Simple (a) and complex (b) hierarchical relationships among correlation clusters

- A simple hierarchy of correlation clusters is exemplified in Figure (a): Two one-dimensional correlation clusters (lines) are embedded in a two-dimensional correlation cluster (a plane).
- A complex hierarchical relationship is given by an intersection of multiple correlation clusters (Figure (b)).

Complex Relationships



(a) (b)

Figure: Results of HiCO on the data sets shown above.

- This embedding can be understood as “multiple inheritance” and, thus, not as a “pure” hierarchy, but as a complex relationship.
- This kind of relationship among correlation clusters confuses a purely hierarchical approach like HiCO.

General Idea

- We would like to find all correlation clusters for all possible correlation dimensions simultaneously.
- We would like to get information concerning relationships (embedding) among correlation clusters of different correlation dimensionality.
- Three steps of algorithm ERiC (Exploring Relationships among Correlation Clusters):
  1. Partition the database objects according to their local correlation dimensionality.
  2. Perform a clustering procedure in each partition (flat clustering, but including information concerning the correlation dimensionality).
  3. Construct a relationship graph bottom up based on the information gathered in step 2.

Partitioning w.r.t. Correlation Dimensionality

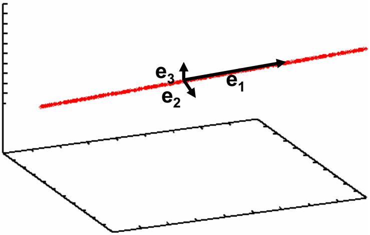
- Basic Assumption: The local neighborhood (e.g. k-NN) of a point (local correlation dimensionality) reflects the correlation dimensionality of a cluster, the point may belong to.
- Thus, for the clustering procedure, we need for a point only to consider those points with equal local correlation dimensionality.
- Having derived the local correlation dimensionality for each point, we partition the database accordingly:
- A point $p \in \mathcal{D}$ with $\lambda_p = i$ is assigned to a partition $\mathcal{D}_i$ of the database $\mathcal{D}$.
- Result: A set of d disjoint subsets $\mathcal{D}_1, \ldots, \mathcal{D}_d$ of $\mathcal{D}$ (some may remain empty).

Gain of Partitioning the Database

By this preprocessing step, we yield several advantages:

- Each point gets assigned an appropriate local correlation dimensionality in advance.
- The number n of data points to process in the following clustering step for each partition is reduced to $\frac{n}{d}$ on the average.
- The clustering procedure can assume all points in a given partition to share a common correlation dimensionality (although not necessarily to belong to a common cluster).

Utilizing the Unified Local Correlation Dimensionality

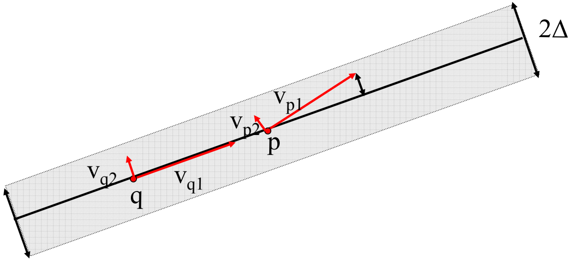


- Each partition of the database can be treated independently in the clustering step.
- For each point, we discern strong and weak eigenvectors.
- Strong eigenvectors span the hyperplane associated with a possible correlation cluster containing the point.
- Weak eigenvectors are perpendicular to this hyperplane.

Approximate Linear Dependency

- Comparing two points p, q, we know that $\lambda_p \leq \lambda_q$ (actually $\lambda_p = \lambda_q$).
- The strong eigenvectors of p are approximately linear dependent from the strong eigenvectors of q, iff for all strong eigenvectors v_i of p:

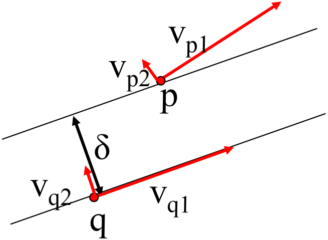
$$\sqrt{v_i^\top \cdot V_q \cdot \hat{E}_q \cdot V_q^\top \cdot v_i} \leq \Delta$$



- Notation: $\textsc{span}(p) \subseteq^{\Delta}_{\text{aff}} \textsc{span}(q)$

Affine Distance

- Two subspaces for the points p and q, $\lambda_p \leq \lambda_q$, may be approximately linear dependent ($\text{SPAN}(p) \subseteq^{\Delta}_{\text{aff}} \text{SPAN}(q)$) but nevertheless p is possibly not in the subspace of q ($p \notin \text{SPAN}(q)$).
- In this case, the subspaces are (approximately) parallel, but not identical.



- The distance between p and q along the weak eigenvectors of q discerns parallel from identical subspaces:

$$\text{DIST}_{\text{aff}}(p, q) = \sqrt{(p - q)^{\mathsf{T}} \cdot V_q \cdot \hat{E}_q \cdot V_q^{\mathsf{T}} \cdot (p - q)}$$

Correlation Distance

Combining *approximate linear dependency* and *affine distance* we yield as correlation distance:

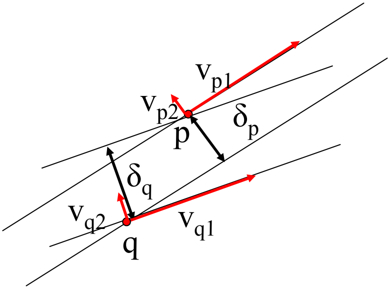
Definition

Let $\delta \in \mathbb{R}_0^+$, $\Delta \in]0, 1[$, $p, q \in \mathcal{D}$, and w.l.o.g. $\lambda_p \leq \lambda_q$. Then the *correlation distance* $\textsc{CorrDist}_\Delta^\delta$ between two points $p, q \in \mathcal{D}$, denoted by $\textsc{CorrDist}_\Delta^\delta(p, q)$, is defined as follows

$$\textsc{CorrDist}_\Delta^\delta(p, q) = \begin{cases} 0 & \text{if } \textsc{span}(p) \subseteq_{\text{aff}}^{\Delta} \textsc{span}(q) \\ & \quad \wedge \textsc{dist}_{\text{aff}}(p, q) \leq \delta \\ 1 & \text{otherwise} \end{cases}$$

Distance Measure for Clustering

- Obviously, the correlation distance is not symmetric:



- Symmetric distance:

$$\text{dist}(p, q) = \max \left(\textsc{CorrDist}_{\Delta}^{\delta}(p, q), \textsc{CorrDist}_{\Delta}^{\delta}(q, p) \right).$$

- Using this distance measure in DBSCAN [Ester et al. (KDD 1996)], we get a set of clusters for each partition $\mathcal{D}_i$ of the database $\mathcal{D}$.

Relationships between Clusters of Different Correlation Dimensionality

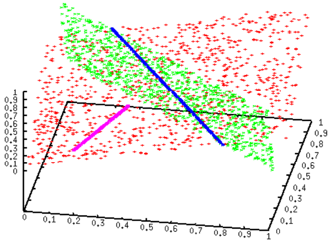
- For aggregating the hierarchical relationships among clusters of different correlation dimensionality, we can employ the definitions above, since we do not need $\lambda_p = \lambda_q$, but only $\lambda_p \leq \lambda_q$.
- Comparing clusters of different correlation dimensionality, $\lambda_p < \lambda_q$ holds.
- Each cluster $\mathcal{C}_i$ is described by its centroid x_i and the set of strong and weak eigenvectors for the centroid w.r.t. all cluster members as neighborhood.

Aggregation of Hierarchical Relationships

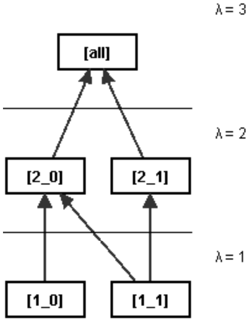
- Assuming the clusters being sorted in ascending order w.r.t. their correlation dimensionality, we start with the first cluster $\mathcal{C}_m$ and check for each cluster $\mathcal{C}_n$ with $\lambda_n > \lambda_m$ whether
  - $\text{SPAN}(x_m) \subseteq^{\Delta}_{\text{aff}} \text{SPAN}(x_n)$ and
  - $\text{DIST}_{\text{aff}}(x_m, x_n) \leq \delta$

  (i.e., $\text{CORRDIST}^{\delta}_{\Delta}(x_m, x_n) = 0$).
- If so, cluster $\mathcal{C}_n$ is treated as parent of cluster $\mathcal{C}_m$, unless $\mathcal{C}_n$ is a parent of any cluster $\mathcal{C}_o$ that in turn is already a parent of $\mathcal{C}_m$.

Visualization of the Resulting Hierarchy



(a) Sample data set 2



(b)

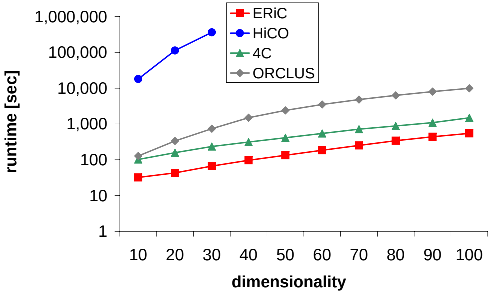
Figure: Example dataset and hierarchical relationship among clusters

Runtime Complexity

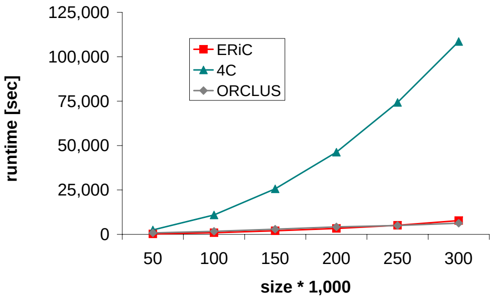
- Preprocessing:
  - k-nearest neighbor query: $O(n)$
  - Based on k-nearest neighbors: $d \times d$ covariance matrix: $O(k \cdot d^2)$
  - Decomposition of covariance matrix (PCA): $O(d^3)$

  For all points: $O(n^2 + k \cdot d^2 \cdot n)$ (since $d << k$)
- Second step (DBSCAN with correlation distance): $O(d^3 \cdot n_i^2)$
  (n_i: number of points in partition i)
  Assuming uniform distribution of the points over all possible correlation dimensionalities: all partitions contain $\frac{n}{d}$ points – for d partitions the runtime reduces to $O(d^2 \cdot n^2)$.
- Aggregation considers all pairs of clusters: $O(|\mathcal{C}|^2 \cdot d^3)$
  Due to $|\mathcal{C}| << n$, the complexity is dominated by the second step: $O(n^2 \cdot d^2)$

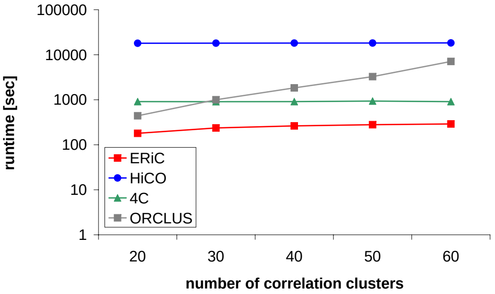
Scalability w.r.t. Dimensionality



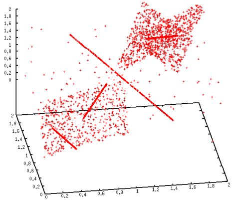
Scalability w.r.t. Size of the Data Set



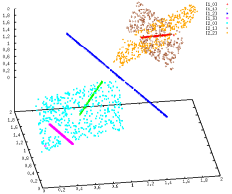
Scalability w.r.t. Number of Clusters



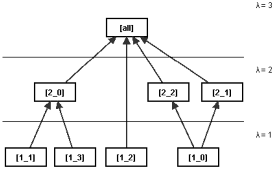
Synthetic Example



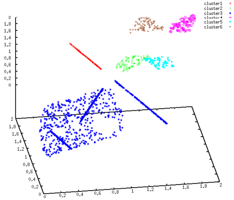
(a) Data set DS1



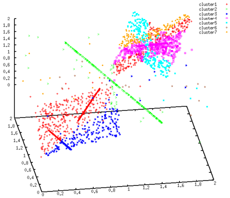
(b) Clusters found by ERiC



(c) Hierarchy generated by ERiC



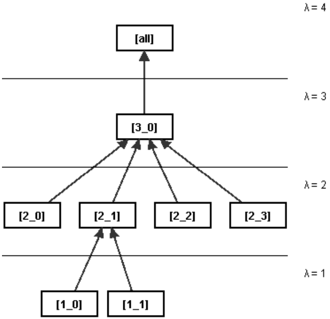
(d) Clusters found by 4C



(e) Clusters found by ORCLUS

(f) Hierarchy generated by HiCO

Wages Data Set



(a) Hierarchy generated by ERiC

cluster	description
1_0	YE = 12, A = 22, YW = 4
1_1	YE = 12, A = 22, YW = 20
2_0	YE = 14, A = YW + 20
2_1	YE = 12, A = YW+18
2_2	YE = 16, A = YW + 22
2_3	YE = 13, A = YW+19
3_0	YE = A - YW - 6

(b) Contents of found clusters

Figure: Results of ERiC on the wages data set.

Pendigits Data Set

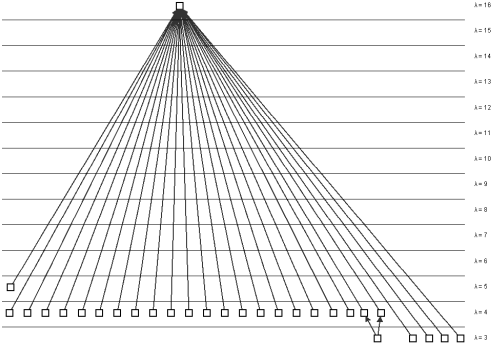


Figure: Hierarchy generated by ERiC on pendigits data set.

Conclusions

- Motivation: Search for complex hierarchies of correlation clusters
- Complex task, state-of-the-art approaches fail in many cases to detect an appropriate clustering structure
- ERiC outperforms the competitors in terms of efficiency and effectivity
- Clear visualization of the cluster hierarchy

References

- E. Achtert, C. Böhm, P. Kröger, and A. Zimek.
  Mining hierarchies of correlation clusters.
  In *Proc. SSDBM*, 2006.
- C. C. Aggarwal and P. S. Yu.
  Finding generalized projected clusters in high dimensional space.
  In *Proc. SIGMOD*, 2000.
- C. Böhm, K. Kailing, P. Kröger, and A. Zimek.
  Computing clusters of correlation connected objects.
  In *Proc. SIGMOD*, 2004.
- M. Ester, H.-P. Kriegel, J. Sander, and X. Xu.
  A density-based algorithm for discovering clusters in large spatial databases with noise.
  In *Proc. KDD*, 1996.

Local Correlation Dimensionality

Definition

Let $\alpha \in]0,1[$, $p \in \mathcal{D}$, and let $\mathcal{N}_p$ denote the set of points in the local neighborhood of p. Then the *local correlation dimensionality* λ_p of the point p is the smallest number of eigenvalues e_i in the eigenvalue matrix $E_{\mathcal{N}_p}$ explaining a portion of at least α of the total variance, i.e.

$$\lambda_p = \min_{r \in \{1,\dots,d\}} \left\{ r \,\middle|\, \frac{\sum_{i=1}^{r} e_i}{\sum_{i=1}^{d} e_i} \geq \alpha \right\}$$

Strong and Weak Eigenvectors

Definition

Let $p \in \mathcal{D}$, λ_p be the local correlation dimensionality of p, and let V_p be the corresponding eigenvectors of the point p based on the local neighborhood $\mathcal{N}_p$ of p. We call the first λ_p eigenvectors of V_p *strong eigenvectors*, the remaining eigenvectors are called *weak*.

Selection Matrix for Weak Eigenvectors

Definition

Let $p \in \mathcal{D}$, λ_p be the local correlation dimensionality of p, and let E_p be the corresponding eigenvectors and eigenvalues of the point p based on the local neighborhood $\mathcal{N}_p$ of p. The *selection matrix* $\hat{E}_p$ *for weak eigenvectors* with entries $\hat{e}_{ij} \in \{0, 1\}$, $i, j = 1, \ldots, d$, is constructed according to the following rule:

$$\hat{e}_{ij} = \begin{cases} 1 & \textit{if} \quad i = j > \lambda_p \\ 0 & \textit{otherwise} \end{cases}$$

Approximate Linear Dependency

Definition

Let $\Delta \in]0, 1[$, $p, q \in \mathcal{D}$, and w.l.o.g. $\lambda_p \leq \lambda_q$. Then the strong eigenvectors of p are *approximately linear dependent* from the strong eigenvectors of q if the following condition holds for all strong eigenvectors v_i of p:

$$\sqrt{v_i^{\mathsf{T}} \cdot V_q \cdot \hat{E}_q \cdot V_q^{\mathsf{T}} \cdot v_i} \leq \Delta$$

If the strong eigenvectors of p are *approximately linear dependent* from the strong eigenvectors of q, we write

$$\text{SPAN}(p) \subseteq^{\Delta}_{\text{aff}} \text{SPAN}(q)$$

Affine Distance

Definition

Let $p, q \in \mathcal{D}$, w.l.o.g. $\lambda_p \leq \lambda_q$, and $\text{SPAN}(p) \subseteq^{\Delta}_{\text{aff}} \text{SPAN}(q)$. The *affine distance* between p and q is given by

$$\text{DIST}_{\text{aff}}(p,q) = \sqrt{(p-q)^{\mathsf{T}} \cdot V_q \cdot \hat{E}_q \cdot V_q^{\mathsf{T}} \cdot (p-q)}$$

Definition

Let $\delta \in \mathbb{R}_0^+$, $\Delta \in]0,1[$, $p, q \in \mathcal{D}$, and w.l.o.g. $\lambda_p \leq \lambda_q$. Then the *correlation distance* $\text{CorrDist}_\Delta^\delta$ between two points $p, q \in \mathcal{D}$, denoted by $\text{CorrDist}_\Delta^\delta(p, q)$, is defined as follows

$$\text{CorrDist}_\Delta^\delta(p,q) = \begin{cases} 0 & \text{if } \text{SPAN}(p) \subseteq_{\text{aff}}^{\Delta} \text{SPAN}(q) \\ & \quad \wedge \text{DIST}_{\text{aff}}(p,q) \leq \delta \\ 1 & \text{otherwise} \end{cases}$$