Mining Hierarchies of Correlation Clusters

Elke Achtert  Christian Böhm  Peer Kröger  Arthur Zimek

Institute for Computer Science
Ludwig-Maximilians-Universität München

10th International Conference on Scientific and Statistical Database Management, Vienna, Austria, 2006
Overview

1. What are Correlation Clusters?
   - Appearance of Correlation Clusters
   - Description of Correlation Clusters

2. Hierarchical Approach to Correlation Clustering
   - Hierarchical Clustering
   - Hierarchical Correlation Clustering

3. Evaluation
   - Synthetic Data
   - Real World Data

4. Conclusions
Correlation Clusters

- Strong correlations between different features may correspond to approximate linear dependencies.
- They appear in the data space as hyperplanes exhibiting a high density of data points.
What are Correlation Clusters?

Description of Correlation Clusters

Covering Correlation Clusters

- derive the **local covariance matrix** $\Sigma_P$ for the $k$-nearest neighbors of a point $P$
- decomposition of $\Sigma_P$ to Eigenvalues and Eigenvectors
- most of the variance covered by small number of Eigenvectors
- number of Eigenvectors covering most of the variance is called **local correlation dimensionality** of a point $P$: $\lambda_P$
- Eigenvectors $\#1 \ldots \#\lambda_P$: **strong** Eigenvectors
- Eigenvectors $\#\lambda_P + 1 \ldots \#d$: **weak** Eigenvectors
Strong and Weak Eigenvectors

- Strong Eigenvectors span the hyperplane corresponding to a correlation cluster.
- Weak Eigenvectors are orthogonal to the hyperplane.

(a) $\lambda = 1$

(b) $\lambda = 2$
General Strategy for Hierarchical Clustering

- keep two separate sets of points
  - points already placed in cluster structure
  - points not yet placed in cluster structure
- each step: select one point of the latter set and place it in the first set
- selection: minimize the distance to any of the points in the first set
General Strategy for Hierarchical Clustering

- keep two separate sets of points
  - points already placed in cluster structure
  - points not yet placed in cluster structure
- each step: select one point of the latter set and place it in the first set
- selection: minimize the distance to any of the points in the first set
General Strategy for Hierarchical Clustering

- keep two separate sets of points
  - points already placed in cluster structure
  - points not yet placed in cluster structure

- each step: select one point of the latter set and place it in the first set

- selection: minimize the distance to any of the points in the first set
General Strategy for Hierarchical Clustering

- keep two separate sets of points
  - points already placed in cluster structure
  - points not yet placed in cluster structure
- each step: select one point of the latter set and place it in the first set
- selection: minimize the distance to any of the points in the first set
General Strategy for Hierarchical Clustering

- keep two separate sets of points
  - points already placed in cluster structure
  - points not yet placed in cluster structure
- each step: select one point of the latter set and place it in the first set
- selection: minimize the distance to any of the points in the first set
Hierarchical Correlation Clusters

- hierarchies of clusters: clusters nested into each other
- e.g. correlation hierarchy: lines nested into planes etc.
- general idea: special distance measure
  correlation distance
    - many attributes highly correlated $\rightarrow$ small value
    - only few attributes highly correlated $\rightarrow$ high value
- strategy: merge points with small correlation distances into
  common clusters
If the strong Eigenvectors of two points together form a line (plane, etc.), they get assigned a correlation distance of 1 (2, etc.).

The distance measure between two points corresponds to the dimensionality of the space spanned by the strong Eigenvectors of the two points.

weaken the algebraic sense of spanning a space to account for slight deviations of a hyperplane.
If the strong Eigenvectors of two points together form a line (plane, etc.), they get assigned a correlation distance of 1 (2, etc.).

The distance measure between two points corresponds to the dimensionality of the space spanned by the strong Eigenvectors of the two points.

Weaken the algebraic sense of spanning a space to account for slight deviations of a hyperplane.
If the **strong** Eigenvectors of two points together form a line (plane, etc.), they get assigned a **correlation distance** of 1 (2, etc.).

The distance measure between two points corresponds to the dimensionality of the space spanned by the strong Eigenvectors of the two points.

weaken the algebraic sense of **spanning a space** to account for slight deviations of a hyperplane.
“Spanning a Space”

- Let a vector $q$ add a new dimension to the space spanned by \{${p_1, \ldots, p_n}$\} if the “difference” between $q$ and this space is substantial, i.e. if it exceeds the threshold parameter $\Delta$
- “difference”: deviation along weak Eigenvectors
- Build local correlation similarity matrix $\hat{M}$ from weak Eigenvectors
Test for “Linear Indepency”

- Test \( q_1 \) for linear independency (in our relaxed sense) to all the strong Eigenvectors \( p_i \) of \( P \):

\[
q_1^T \cdot \hat{M}_P \cdot q_1 > \Delta^2
\]

- If so, \( q_1 \) opens up a new dimension compared to \( P \). The correlation dimensionality \( \lambda(Q, P) \) is at least \( \lambda_P + 1 \).

- Test a second vector \( q_2 \):
  Is \( q_2 \) “linearly independent” from strong Eigenvectors of \( P \cup q_1 \)?

...
Test for “Linear Independency”

- Test $q_1$ for linear independency (in our relaxed sense) to all the strong Eigenvectors $p_i$ of $P$:

$$q_1^T \hat{M}_P q_1 > \Delta^2$$

- If so, $q_1$ opens up a new dimension compared to $P$. The correlation dimensionality $\lambda(Q, P)$ is at least $\lambda_P + 1$.

- Test a second vector $q_2$:
  - Is $q_2$ “linearly independent” from strong Eigenvectors of $P \cup q_1$?
  - ...
Test for “Linear Independency”

- Test \( q_1 \) for linear independency (in our relaxed sense) to all the strong Eigenvectors \( p_i \) of \( P \):
  \[
  q_1^T \cdot \hat{M}_P \cdot q_1 > \Delta^2
  \]

  If so, \( q_1 \) opens up a new dimension compared to \( P \). The correlation dimensionality \( \lambda(Q, P) \) is at least \( \lambda_P + 1 \).

- Test a second vector \( q_2 \):
  Is \( q_2 \) “linearly independent” from strong Eigenvectors of \( P \cup q_1 \)?

  ...
Formalization of the Correlation Distance

Definition

The correlation distance between two points $P, Q \in \mathcal{D}$, denoted by $\text{CDIST}(P, Q)$, is a pair consisting of the correlation dimensionality of $P$ and $Q$ and the Euclidean distance between $P$ and $Q$, i.e.

$$\text{CDIST}(P, Q) = (\lambda(P, Q), \text{dist}(P, Q)).$$

We say $\text{CDIST}(P, Q) \leq \text{CDIST}(R, S)$ if one of the following conditions holds:

1. $\lambda(P, Q) < \lambda(R, S)$,
2. $\lambda(P, Q) = \lambda(R, S) \land \text{dist}(P, Q) \leq \text{dist}(R, S)$. 
Given the correlation distance measure, any hierarchical clustering algorithm based on distance comparisons could be employed to seek for correlation cluster hierarchies.

We used the algorithmic schema of OPTICS.

Our approach: HiCO (Hierarchical Correlation Ordering)

Like OPTICS, HiCO visualizes the cluster hierarchy in a cluster-order as a plot of the so called reachability distances.
Algorithmic Schema and Result Representation

1. Select a not yet processed object \( o \).
2. Add \( o \) to \( CO \) and determine neighbors of \( o \).
3. Determine reachability distance of each neighbor.
4. Remove the first object \( o \) from the seed list.
5. If the seed list is not empty, insert or update the neighbors in the seed list.
6. If the seed list is empty, stop the process.
Synthetic Data Set
HiCO - Cluster Order
HiCO - Cluster Order

(a) Cluster 1  (b) Cluster 2  (c) Cluster 3
Exemplary Results: Metabolome Data
“Correlation Clusters” are clusters of points exhibiting possible linear dependencies among several features.

The hierarchical clustering approach enables us to find clusters in different ranges simultaneously.

We introduced a correlation distance measure to account for different ranges of correlation dimensionality.

In contrast to existing work, HiCO does not require the user to specify

- any global density threshold,
- the number of clusters to be found,
- nor any parameter specifying the dimensionality of the clusters.

Results show HiCO finding meaningful correlation clusters of lower dimensionality embedded in correlation clusters of higher dimensionality, superior to other approaches.
Other Approaches

- Subspace (Projected) Clustering: finds axis parallel projections only
- Pattern-Based Clustering (aka. Co-Clustering or Bi-Clustering): limited to pairwise positive correlations
- Correlation Clustering:
  - **ORCLUS**: integrates PCA into $k$-means — user needs to specify number of clusters in advance
  - **4C**: integrates PCA into DBSCAN — user needs to specify global density threshold

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

(a) Cluster 1     (b) Cluster 2     (c) Cluster 3
Appendix
Results of Other Methods

OPTICS

(a) Reachability Plot
(b) Cluster 1
(c) Cluster 2

Achtert, Böhm, Kröger, Zimek (LMU)
Mining Hierarchies of Correlation Clusters
SSDBM 2006
Appendix

Results of Other Methods

4C

(a) $\lambda = 1$

(b) $\lambda = 2$

Achtert, Böhm, Kröger, Zimek (LMU) Mining Hierarchies of Correlation Clusters SSDBM 2006
Local Covariance Matrix

Definition

Let \( k \in \mathbb{N} \), \( k \leq |\mathcal{D}| \). The local covariance matrix \( \Sigma_P \) of a point \( P \in \mathcal{D} \) w.r.t. \( k \) is formed by the \( k \) nearest neighbors of \( P \). Let \( \overline{X} \) be the centroid of \( \text{NN}_k(P) \), then

\[
\Sigma_P = \frac{1}{|\text{NN}_k(P)|} \sum_{X \in \text{NN}_k(P)} (X - \overline{X}) \cdot (X - \overline{X})^T
\]

Since the local covariance matrix \( \Sigma_P \) of a point \( P \) is a square matrix it can be decomposed into the Eigenvalue matrix \( E_P \) of \( P \) and the Eigenvector matrix \( V_P \) of \( P \) such that \( \Sigma_P = V_P \cdot E_P \cdot V_P^T \).
Local Correlation Similarity Matrix

**Definition**

Let point \( P \in \mathcal{D} \), \( V_P \) the corresponding \( d \times d \) Eigenvector matrix of the local covariance matrix \( \Sigma_P \) of \( P \), and \( \lambda_P \) the local correlation dimensionality of \( P \). The matrix \( \hat{E}_P \) with entries \( \hat{e}_i (i = 1, \ldots, d) \) is computed according to the following rule:

\[
\hat{e}_i = \begin{cases} 
0, & \text{if } i \leq \lambda_P \\
1, & \text{otherwise}
\end{cases}
\]

The matrix

\[
\hat{M}_P = V_P \hat{E}_P V_P^T
\]

is called the **local correlation similarity matrix** of \( P \).
The local correlation similarity matrix is suitable to define a quadratic form distance measure w.r.t. a point:

**Definition**

The **local correlation distance** of point $P$ to point $Q$ according to the local correlation similarity matrix $\hat{M}_P$ associated with point $P$ is denoted by

$$\text{LOCDIST}_P(P, Q) = \sqrt{(P - Q)^T \cdot \hat{M}_P \cdot (P - Q)}.$$
Effect of the Local Correlation Distance

- Weights distances along the strong Eigenvectors by 0.
- Weights distances along the weak Eigenvectors by 1.
- Only distances orthogonal to the cluster hyperplane are relevant.
Example: Metabolic Pathways

- There are certain pathways for degradation of metabolics.
- Concentrations of input and output metabolites may be correlated, the concentration of alternative intermediate states may vary depending on the environment.
- Genetic disorders may lead to failure of some pathways, other pathways are used more intensely.
- The concentrations of more metabolites are correlated if samples suffer from certain diseases.
Example: Metabolic Pathways

- There are certain pathways for degradation of metabolics.
- Concentrations of input and output metabolites may be correlated, the concentration of alternative intermediate states may vary depending on the environment.
- Genetic disorders may lead to failure of some pathways, other pathways are used more intensely.
- The concentrations of more metabolites are correlated if samples suffer from certain diseases.
Example: Metabolic Pathways

- There are certain pathways for degradation of metabolics.
- Concentrations of input and output metabolites may be correlated, the concentration of alternative intermediate states may vary depending on the environment.
- Genetic disorders may lead to failure of some pathways, other pathways are used more intensely.
- The concentrations of more metabolites are correlated if samples suffer from certain diseases.
Example: Metabolic Pathways

- There are certain pathways for degradation of metabolics.
- Concentrations of input and output metabolites may be correlated, the concentration of alternative intermediate states may vary depending on the environment.
- Genetic disorders may lead to failure of some pathways, other pathways are used more intensely.
- The concentrations of more metabolites are correlated if samples suffer from certain diseases.
Correlation Dimensionality

The correlation dimensionality between two points $P, Q \in D$, denoted by $\lambda(P, Q)$, is the dimensionality of the space which is spanned by the union of the strong Eigenvectors associated to $P$ and the strong Eigenvectors associated to $Q$.

All four vectors are pairwise linearly independent. But the union of all four is spanning a space of dimensionality 3.
Considerations for the Correlation Distance

- The dimensionality of the spaces spanned by unifying the strong Eigenvectors of \( P \) with the set of strong Eigenvectors of \( Q \) or vice versa can differ from each other, i.e. \( \lambda_P(P, Q) \) and \( \lambda_Q(P, Q) \) may differ.

- As a symmetric distance measure we build the maximum:

  \[
  \lambda(P, Q) = \max (\lambda_P(P, Q), \lambda_Q(P, Q))
  \]

- As \( \lambda(P, Q) \in \mathbb{N} \), many distances between different point pairs are identical. Resolve tie situations by additionally considering the Euclidean distance.

- As a consequence, inside a correlation cluster the points are clustered as by a conventional hierarchical clustering method.