Data Mining and the 'Curse of Dimensionality'

iDB Workshop 2011

Arthur Zimek

Ludwig-Maximilians-Universität München
Munich, Germany
http://www.dbs.ifi.lmu.de/~zimek
zimek@dbs.ifi.lmu.de
Outline

1. The Curse of Dimensionality
2. Shared-Neighbor Distances
3. Subspace Outlier Detection
4. Subspace Clustering
5. Conclusions
The “curse of dimensionality”: one buzzword for many problems [KKZ09]

• First aspect: *Optimization Problem* (Bellman).
  
  “[The] curse of dimensionality […] is a malediction that has plagued the scientists from earliest days.” [Bel61]
  
  – The difficulty of any global optimization approach increases exponentially with an increasing number of variables (dimensions).
  
  – General relation to clustering: fitting of functions (each function explaining one cluster) becomes more difficult with more degrees of freedom.
  
  – Direct relation to subspace clustering: number of possible subspaces increases dramatically with increasing number of dimensions.
The Curse of Dimensionality

- Second aspect: *Concentration effect of $L_p$-norms*
  - In [BGRS99,HAK00] it is reported that the ratio of $(D_{max} - D_{min})$ to $D_{min}$ converges to zero with increasing dimensionality $d$
    - $D_{min} = \text{distance to the nearest neighbor in } d \text{ dimensions}$
    - $D_{max} = \text{distance to the farthest neighbor in } d \text{ dimensions}$
  
  Formally:

  $\forall \epsilon > 0 : \lim_{d \to \infty} P\left[ \text{dist}_d\left( \frac{D_{max} - D_{min}}{D_{min}}, 0 \right) \leq \epsilon \right] = 1$

  - Distances to near and to far neighbors become more and more similar with increasing data dimensionality (loss of *relative contrast* or *concentration effect* of distances).
  - This holds true for a wide range of data distributions and distance functions, but…
The Curse of Dimensionality

From bottom to top: minimum observed value, average minus standard deviation, average value, average plus standard deviation, maximum observed value, and maximum possible value of the Euclidean norm of a random vector. The expectation grows, but the variance remains constant. A small subinterval of the domain of the norm is reached in practice. (Figure and caption: [FWV07])

- The observations stated in [BGRS99, HAK00, AHK01] are valid within clusters but not between different clusters as long as the clusters are well separated [BFG99, FWV07, HKK+10].
- This is not the main problem for subspace clustering, although it should be kept in mind for range queries.
• Third aspect: *Relevant and Irrelevant attributes*
  – A subset of the features may be relevant for clustering
  – Groups of similar (“dense”) points may be identified when considering these features only
  – Different subsets of attributes may be relevant for different clusters
  – Separation of clusters relates to *relevant attributes* (helpful to discern between clusters) as opposed to *irrelevant attributes* (indistinguishable distribution of attribute values for different clusters).
Effect on clustering:
- Usually the distance functions used give equal weight to all dimensions.
- However, not all dimensions are of equal importance.
- Adding irrelevant dimensions ruins any clustering based on a distance function that equally weights all dimensions.
The Curse of Dimensionality

Zimek: Data Mining and the 'Curse of Dimensionality' (IDB Workshop 2011)
The Curse of Dimensionality

• Fourth aspect: *Correlation among attributes (redundancy?)*
  – A subset of features may be correlated
  – Groups of similar (“dense”) points may be identified when considering this correlation of features only
  
  – different correlations of attributes may be relevant for different clusters
  – can result in lower intrinsic dimensionality of a data set
  – bad discrimination of distances can still be a problem

Zimek: Data Mining and the 'Curse of Dimensionality' (IDB Workshop 2011)
• there are other effects of the “curse of dimensionality”
• just another strange fact: the volume of hyperspheres shrinks with increasing dimensionality!

\[ V_n(r) = \frac{\pi^{\frac{n}{2}} r^n}{\Gamma\left(\frac{n}{2} + 1\right)} \]
[HKK+10]: Can Shared-Neighbor Distances Defeat the Curse of Dimensionality? (SSDBM 2010)

• we mainly aim at distinguishing these effects of the ’curse’:
  – concentration effect within distributions
  – impediment of similarity search by irrelevant attributes
  – partly: impact of redundant/correlated attributes

• as a remedy for similarity assessment in high dimensional data, to use shared nearest neighbor (SNN) information has been proposed but never evaluated systematically

• [HKK+10]: evaluation of the effects on primary distances (Manhattan, Euclidean, fractional $L_p$ ($L_{0.6}$ and $L_{0.8}$), cosine) and secondary distances (SNN)
Shared-Neighbor Distances

- secondary distances are defined on top of primary distances
- shared nearest neighbor (SNN) information:
  - assess the set of $s$ nearest neighbors for two objects $x$ and $y$ in terms of some primary distance (Euclidean, Manhattan, cosine…)
  - derive overlap of neighbors (common objects in the NN of $x$ and $y$)

$$SNN_s(x, y) = |NN_s(x) \cap NN_s(y)|$$

- similarity measure

$$\text{simcos}_s(x, y) = \frac{SNN_s(x, y)}{s}$$

- cosine of the angle between membership vectors for NN($x$) and NN($y$)

- SNN has been used before in mining high-dimensional data, but alleged quality improvement has never been evaluated
Shared-Neighbor Distances

- distance measures based on SNN:
  \[ d_{\text{inv}}(x, y) = 1 - \text{simcos}_s(x, y) \]
  \[ d_{\text{acos}}(x, y) = \arccos(\text{simcos}_s(x, y)) \]
  \[ d_{\ln}(x, y) = -\ln(\text{simcos}_s(x, y)) \]
  - \( d_{\text{inv}} \): linear inversion
  - \( d_{\text{acos}} \) penalizes slightly suboptimal similarities more strongly
  - \( d_{\ln} \) more tolerant for relatively high similarity values but approaches infinity for very low similarity values

- for assessment of ranking quality, these formulations are equivalent as the ranking is unaffected

- only \( d_{\text{acos}} \) is a metric (if the underlying primary distance is a metric)
Shared-Neighbor Distances

- Artificial data sets: $n = 10,000$ items, $c = 100$ clusters, up to $d = 640$ dimensions, cluster sizes randomly determined.
- Relevant attribute values normally distributed, irrelevant attribute values uniformly distributed.
- Data sets:
  - All-Relevant: all dimensions relevant for all clusters
  - 10-Relevant: first 10 dimensions are relevant for all clusters, the remaining dimensions are irrelevant
  - Cyc-Relevant: $i$th attribute is relevant for the $j$th cluster when $i \mod c = j$, otherwise irrelevant (here: $c = 10$, $n = 1000$)
  - Half-Relevant: for each cluster, an attribute is chosen to be relevant with probability 0.5, and irrelevant otherwise
  - All-Dependent: derived from All-Relevant introducing correlations among attributes $X \in AllDependent, Y \in AllRelevant: X_i = Y_i \ (1 \leq i \leq 10), \ X_i = \frac{1}{2} (X_{i-10} + Y_i) \ (i > 10)$
  - 10-Dependent: derived from 10-Relevant introducing correlations among attributes
Data sets show properties of the “curse of dimensionality”

\[
\lim_{d \to \infty} \frac{\text{var} \left( \frac{\|X_d\|}{E\|X_d\|} \right)}{D_{\text{max}} - D_{\text{min}}} = 0 \quad \Rightarrow \quad \frac{D_{\text{max}} - D_{\text{min}}}{D_{\text{min}}} \to 0
\]

Zimek: Data Mining and the ‘Curse of Dimensionality’ (iDB Workshop 2011)
Shared-Neighbor Distances

Data sets show properties of the “curse of dimensionality”

\[
\lim_{d \to \infty} \text{var} \left( \frac{\left\| X_d \right\|}{E \left\| X_d \right\|} \right) = 0 \quad \Rightarrow \quad \frac{D_{\text{max}} - D_{\text{min}}}{D_{\text{min}}} \to 0
\]

Zimek: Data Mining and the ‘Curse of Dimensionality’ (IDB Workshop 2011)
Shared-Neighbor Distances

- Using each item in turn as a query, neighborhood ranking reported in terms of the Area under curve (AUC) of the Receiver Operating Characteristic (ROC)

Zimek: Data Mining and the ‘Curse of Dimensionality’ (IDB Workshop 2011)
Shared-Neighbor Distances

Euclidean distance

Zimek: Data Mining and the ‘Curse of Dimensionality’ (iDB Workshop 2011)
Shared-Neighbor Distances

SNN based on Euclidean

All-Relevant
20/40/80/160/320/640 dimensions

Zimek: Data Mining and the 'Curse of Dimensionality' (iDB Workshop 2011)
Shared-Neighbor Distances

SNN based on Euclidean

10-Relevant
20/40/80/160/320/640 dimensions

Zimek: Data Mining and the 'Curse of Dimensionality' (iDB Workshop 2011)
Shared-Neighbor Distances

some real data sets: distributions of Euclidean distances

Zimek: Data Mining and the 'Curse of Dimensionality' (iDB Workshop 2011)
Shared-Neighbor Distances

some real data sets: distributions of SNN distances (Euclidean)

Zimek: Data Mining and the 'Curse of Dimensionality' (iDB Workshop 2011)
Shared-Neighbor Distances

some real data sets: ranking quality

Zimek: Data Mining and the ‘Curse of Dimensionality’ (iDB Workshop 2011)
Subspace Outlier Detection

[KKSZ09]: Outlier Detection in Axis-Parallel Subspaces of High Dimensional Data (PAKDD 2009)

general idea:
• assign a set of reference points to a point \( o \)
  (e.g., \( k \)-nearest neighbors – but keep in mind the “curse of dimensionality”: local feature relevance vs. meaningful distances)
• find the subspace spanned by these reference points (allowing some jitter)
• analyze for the point \( o \) how well it fits to this subspace

Zimek: Data Mining and the ‘Curse of Dimensionality’ (iDB Workshop 2011)
Subspace Outlier Detection

- distance of $o$ to the reference hyperplane:

$$
\text{dist}(o, H(S)) = \sqrt{\sum_{i=1}^{d} \nu_i \cdot (o_i - \mu_i^S)^2}
$$

- the higher this distance, the more deviates the point $o$ from the behavior of the reference set, the more likely it is an outlier

Zimek: Data Mining and the 'Curse of Dimensionality' (IDB Workshop 2011)
Subspace Outlier Detection

subspace outlier degree (SOD) of a point $p$: 

$$SOD_{R(p)}(p) = \frac{dist(o, H(R(p)))}{\nu^{R(p)}}$$

i.e., the distance normalized by the number of contributing attributes

possible normalization to a probability-value $[0,1]$ in relation to the distribution of distances of all points in $S$
Choice of a reference set for outliers?

• recall “curse of dimensionality”
  – local feature relevance → need for a local reference set
  – distances loose expressiveness → how to choose a meaningful local reference set?

• consider k nearest neighbors in terms of the shared nearest neighbor similarity
  – given a primary distance function \( \text{dist} \) (e.g. Euclidean distance)
  – \( N_k(p) \): \( k \)-nearest neighbors in terms of \( \text{dist} \)
  – SNN similarity for two points \( p \) and \( q \):
    \[
    \text{sim}_{\text{SNN}}(p, q) = |N_k(p) \cap N_k(q)|
    \]
  – reference set \( R(p) \): \( l \)-nearest neighbors of \( p \) using \( \text{sim}_{\text{SNN}} \)

Zimek: Data Mining and the ‘Curse of Dimensionality’ (IDB Workshop 2011)
Subspace Outlier Detection

2-d sample data, comparison to

- LOF [BKNS00]
- ABOD [KSZ08]

Zimek: Data Mining and the 'Curse of Dimensionality' (IDB Workshop 2011)
Subspace Outlier Detection

- Gaussian distribution in 3 dimensions, 20 outliers
- adding 7, 17, 27, 47, 67, 97 irrelevant attributes

Zimek: Data Mining and the 'Curse of Dimensionality' (IDB Workshop 2011)
Subspace Clustering

[ABD+08]: *Robust clustering in arbitrarily oriented subspaces* (SDM 2008) (extended version: [ABD+08a])

- Algorithm CASH: Clustering in Arbitrary Subspaces based on the Hough-Transform

- Hough-transform:
  - developed in computer-graphics
  - 2-dimensional (image processing)

- CASH:
  - generalization to $d$-dimensional spaces
  - transfer of the clustering to a new space (“Parameter-space” of the Hough-transform)
  - restriction of the search space (from innumerable infinite to $O(n!)$)
  - common search heuristic for Hough-transform: $O(2^d)$
    $\rightarrow$ efficient search heuristic

Zimek: Data Mining and the ‘Curse of Dimensionality’ (IDB Workshop 2011)
Subspace Clustering

- given: $D \subseteq \mathbb{R}^d$
- find linear subspaces accommodating many points
- Idea: map points from data space (picture space) onto functions in parameter space

Zimek: Data Mining and the 'Curse of Dimensionality' (IDB Workshop 2011)
Subspace Clustering

- $e_i, 1 \leq i \leq d$: orthonormal-basis
- $x = (x_1, \ldots, x_d)^T$: $d$-dimensional vector onto hypersphere around the origin with radius $r$
- $u_i$: unit-vector in direction of projection of $x$ onto subspace $\text{span}(e_i, \ldots, e_d)$
- $\alpha_1, \ldots, \alpha_{d-1}$: $\alpha_i$ angle between $u_i$ and $e_i$

$$x_i = r \cdot \left( \prod_{j=1}^{i-1} \sin(\alpha_j) \right) \cdot \cos(\alpha_i)$$
Subspace Clustering

Length $\delta$ of the normal vector $\delta \cdot \vec{n}$ with $\|\vec{n}\| = 1$ and angles $\alpha_1, \ldots, \alpha_{d-1}$ for the line through point $p$:

$$f_p(\alpha_1, \ldots, \alpha_{d-1}) = \langle p, n \rangle = \sum_{i=1}^{d} p_i \left( \prod_{j=1}^{i-1} \sin(\alpha_j) \right) \cdot \cos(\alpha_i)$$
Subspace Clustering

- Properties of the transformation
  - Point in the data space = sinusoidal curve in parameter space
  - Point in parameter space = hyper-plane in data space
  - Points on a common hyper-plane in data space = sinusoidal curves through a common point in parameter space
  - Intersections of sinusoidal curves in parameter space = hyper-plane through the corresponding points in data space
Subspace Clustering

• dense regions in parameter space \iff linear structures in data space (hyperplanes with $\lambda \leq d-1$)

• exact solution: find all intersection points
  – infeasible
  – too exact

• approximative solution: grid-based clustering in parameter space
  \rightarrow find grid cells intersected by at least $m$ sinusoids
  – search space bounded but in $O(r^d)$
  – pure clusters require large value for $r$ (grid solution)

Zimek: Data Mining and the ‘Curse of Dimensionality’ (IDB Workshop 2011)
Subspace Clustering

efficient search heuristic for dense regions in parameter space

- construct a grid by recursively splitting the parameter space (best-first-search)
- identify dense grid cells as intersected by many parametrization functions
- dense grid cell represents \((d-1)\)-dimensional linear structure
- transform corresponding data objects in corresponding \((d-1)\)-dimensional space and repeat the search recursively

Zimek: Data Mining and the ‘Curse of Dimensionality’ (IDB Workshop 2011)
Subspace Clustering

- grid cell representing less than \( m \) points can be excluded → early pruning of a search path
- grid cell intersected by at least \( m \) sinusoids after \( s \) recursive splits represents a correlation cluster (with \( \lambda \leq d-1 \))
  - remove points of the cluster (and corr. sinusoids) from remaining cells

Zimek: Data Mining and the ‘Curse of Dimensionality’ (IDB Workshop 2011)
Subspace Clustering

- search heuristic: linear in number of points, but $\sim O(d^3)$
  - depth of search $s$, number $c$ of pursued paths (ideally: $c$ cluster):
    - priority search: $O(s \cdot c)$
    - determination of curves intersecting a cell: $O(n \cdot d^3)$
    - overall: $O(s \cdot c \cdot n \cdot d^3)$
      (note: PCA generally in $O(d^3)$)

Zimek: Data Mining and the 'Curse of Dimensionality' (iDB Workshop 2011)
Subspace Clustering

(a) Data set

(b) CASH: Cluster 1-5

(c) 4C: Cluster 1-8

(d) ORCLUS: Cluster 1-5
Subspace Clustering

- stability with increasing number of noise objects
Conclusions

• The *curse of dimensionality* does not count in general as an excuse for everything – depends on the number and nature of distributions in a data set
• the nature of each particular problem needs to be studied in its own
• part of the curse: it’s always different than expected
• if you ever think, you have solved the problems of the curse: watch out for the curse striking back!

Zimek: Data Mining and the ‘Curse of Dimensionality’ (iDB Workshop 2011)
Some Advice

• do not take everything for granted which is stated in the literature
• consider claims in the literature:
  – is there enough evidence to support the claims?
  – is the interpretation of the claims clear?
  – challenge them or support them
• papers report the strengths – you should try to find out the weaknesses and to improve
• have fun!
Robust clustering in arbitrarily oriented subspaces.  
In Proceedings of the 8th SIAM International Conference on Data Mining (SDM),  
Atlanta, GA, 2008

[ABD+08a] E. Achtert, C. Böhm, J. David, P. Kröger, A. Zimek:  
*Global Correlation Clustering Based on the Hough Transform*  

Deriving quantitative models for correlation clusters.  
In Proceedings of the 12th ACM International Conference on Knowledge Discovery  
and Data Mining (SIGKDD), Philadelphia, PA, 2006.

Robust, complete, and efficient correlation clustering.  
In Proceedings of the 7th SIAM International Conference on Data Mining (SDM),  

[AHK01] C. C. Aggarwal, A. Hinneburg, and D. Keim.  
On the surprising behavior of distance metrics in high dimensional space.  
In Proceedings of the 8th International Conference on Database Theory (ICDT),  

Density-based indexing for approximate nearest-neighbor queries.
In Proceedings of the 5th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD), San Diego, CA, 1999.

When is “nearest neighbor” meaningful?
In Proceedings of the 7th International Conference on Database Theory (ICDT), Jerusalem, Israel, 1999.

LOF: Identifying Density-based Local Outliers.

The concentration of fractional distances.
IEEE Transactions on Knowledge and Data Engineering, 19(7): 873-886, 2007.
What is the nearest neighbor in high dimensional spaces? 
In Proceedings of the 26th International Conference on Very Large Data Bases (VLDB), Cairo, Egypt, 2000.

Can Shared-Neighbor Distances Defeat the Curse of Dimensionality? 
In Proceedings of the 22nd International Conference on Scientific and Statistical Data Management (SSDBM), Heidelberg, Germany, 2010.

[KKSZ09] H.-P. Kriegel, P. Kröger, E. Schubert, A. Zimek: 
Outlier Detection in Axis-Parallel Subspaces of High Dimensional Data. 
In Proc. 13th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD), Bangkok, Thailand, 2009.
*Clustering High Dimensional Data: A Survey on Subspace Clustering, Pattern-based Clustering, and Correlation Clustering.*  
ACM Transactions on Knowledge Discovery from Data (TKDD), Volume 3, Issue 1 (March 2009), Article No. 1, pp. 1-58, 2009.

*Angle-Based Outlier Detection in High-dimensional Data.*  
In Proceedings of the 14th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD), Las Vegas, NV, 2008.