Can Shared-Neighbor Distances Defeat the Curse of Dimensionality?

SSDBM 2010

Michael E. Houle[1], Hans-Peter Kriegel[2], Peer Kröger[2], Erich Schubert[2], Arthur Zimek[2]

[1]National Institute of Informatics
Tokyo, Japan
meh@nii.ac.jp

[2]Ludwig-Maximilians-Universität München
Munich, Germany
http://www.dbs.ifi.lmu.de
{kriegel,kroegerp,schube,zimek}@dbs.ifi.lmu.de

Outline

1. The Curse of Dimensionality
2. Shared-Neighbor Distances
3. Experimental Set-Up
4. Observations
5. Conclusions

The Curse of Dimensionality

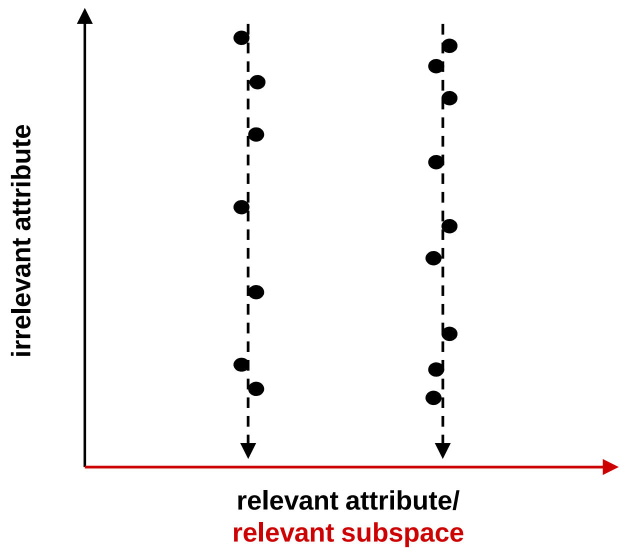
- Beyer et al. (1999): distances to near and to far neighbors become more and more similar with increasing data dimensionality (loss of *relative contrast* or *concentration effect* of distances):

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

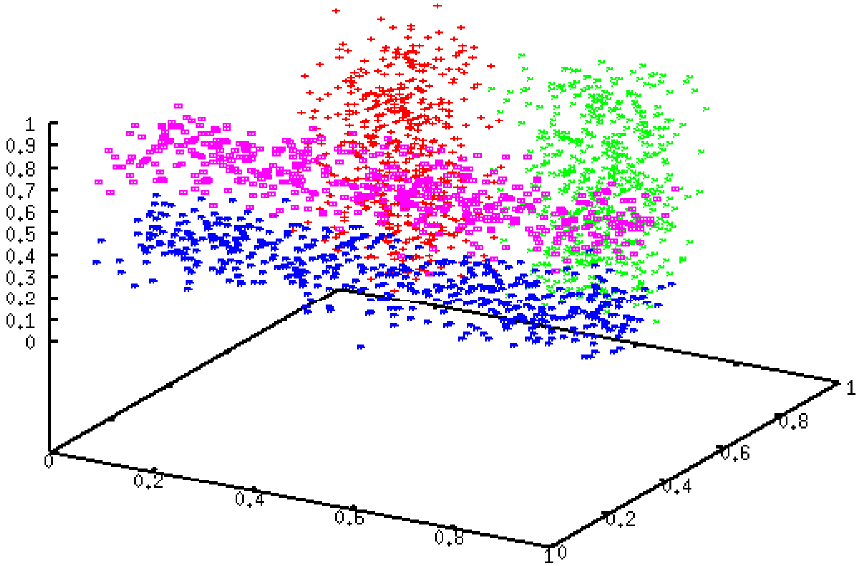
- valid for a broad range of data distributions
- but only within one single distribution

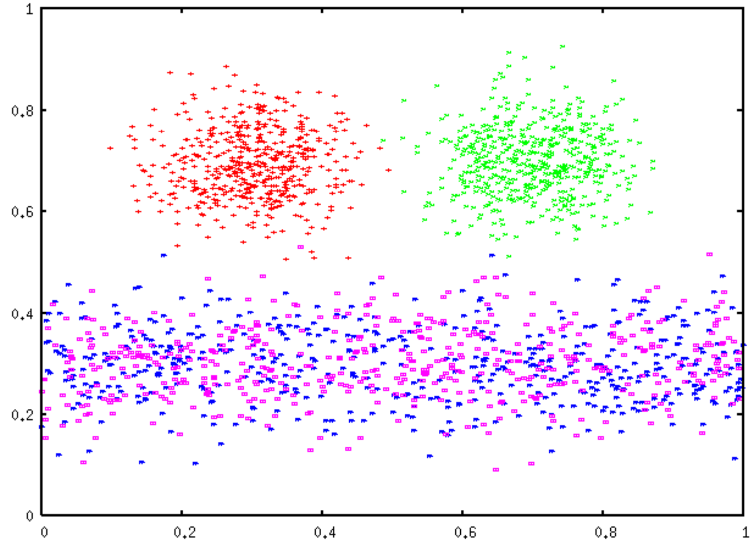
The Curse of Dimensionality

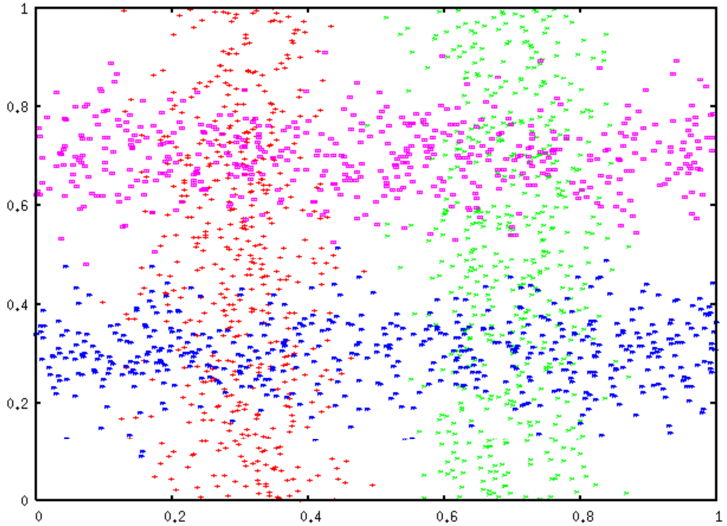
- Bennett et al. (1999): nearest-neighbor queries are still meaningful, if the search is limited to the same cluster and if the clusters are well separated.
- Separation of clusters relates to *relevant attributes* (helpful to discern between clusters) as opposed to *irrelevant attributes* (indistinguishable distribution of attribute values for different clusters).

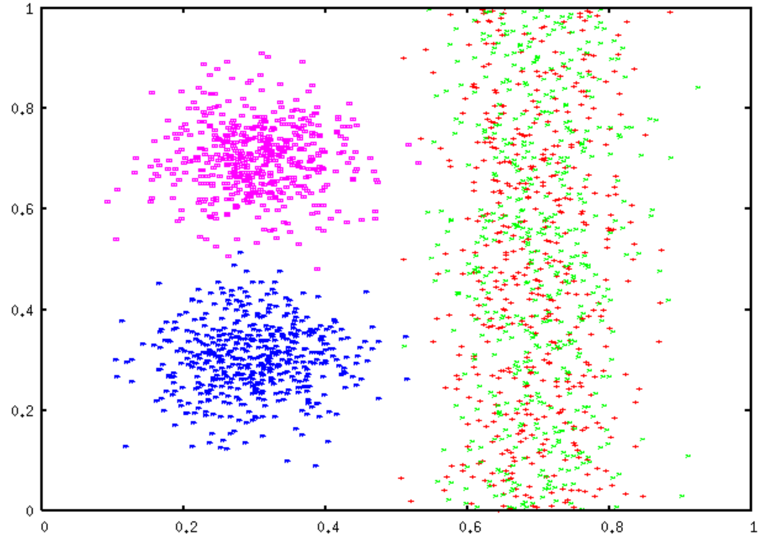


The Curse of Dimensionality



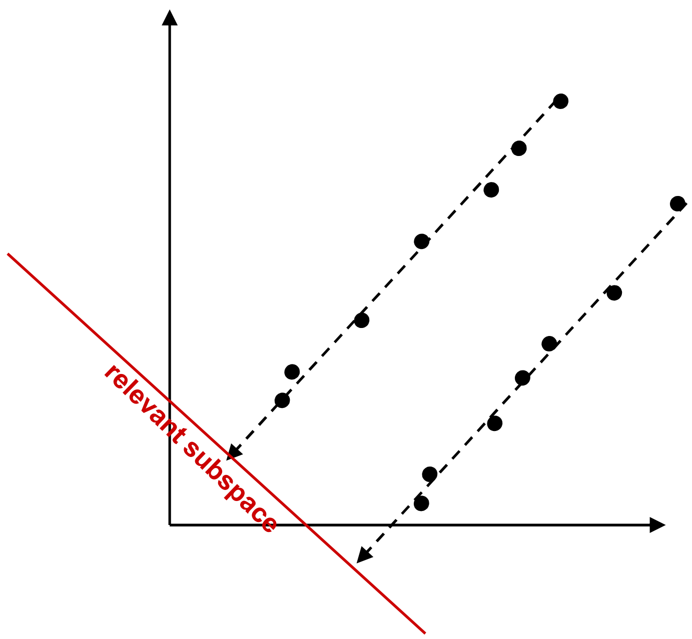






The Curse of Dimensionality

- Redundant attributes: dependencies/correlations among attributes
  - can result in lower intrinsic dimensionality of a data set
  - bad discrimination of distances can still be a problem



The Curse of Dimensionality

- there are other effects of the “curse of dimensionality”
- we mainly aim at distinguishing these effects:
  - 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
- here: 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*)

$$\mathrm{SNN}_s(x,y) = \left|\mathrm{NN}_s(x) \cap \mathrm{NN}_s(y)\right|$$

  - similarity measure

$$\mathrm{simcos}_s(x,y) = \frac{\mathrm{SNN}_s(x,y)}{s}$$

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:

$$\mathrm{dinv}_s(x, y) = 1 - \mathrm{simcos}_s(x, y)$$

$$\mathrm{dacos}_s(x, y) = \arccos(\mathrm{simcos}_s(x, y))$$

$$\mathrm{dln}_s(x, y) = -\ln(\mathrm{simcos}_s(x, y))$$

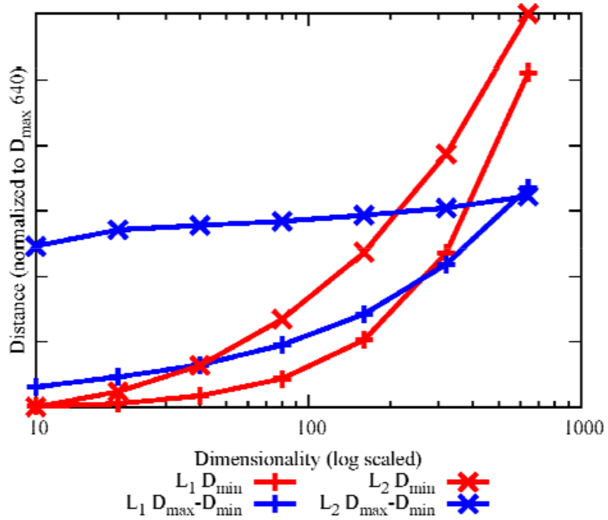
  - dinv: linear inversion
  - dacos penalizes slightly suboptimal similarities more strongly
  - dln more tolerant for relatively high similarity values but approaches infinity for very low similarity values
- for assessment of ranking quality, these formulations are equivalent as the ranking is unaffected
- only dacos is a metric (if the underlying primary distance is a metric)

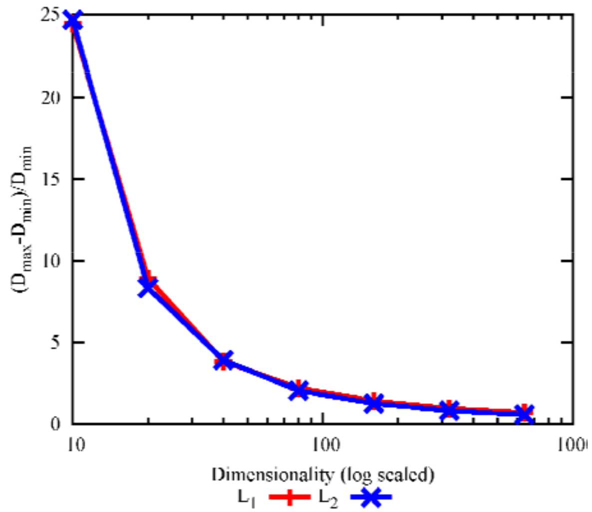
Experimental Set-Up

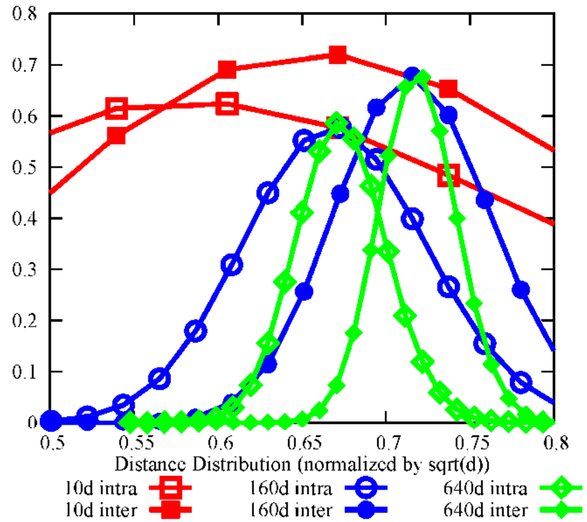
- Artificial data sets: *n = 10.000* items, *c = 100* clusters, up to *d = 640* dimensions, cluster sizes randomly determined.
- Relevant attribute values normally distributed, irrelevant attribute values uniformly distributed.
- Data sets:
  - *All-Relevant:* all dimensions relevant for all clusters
  - *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

Experimental Set-Up

Data sets show properties of the “curse of dimensionality”





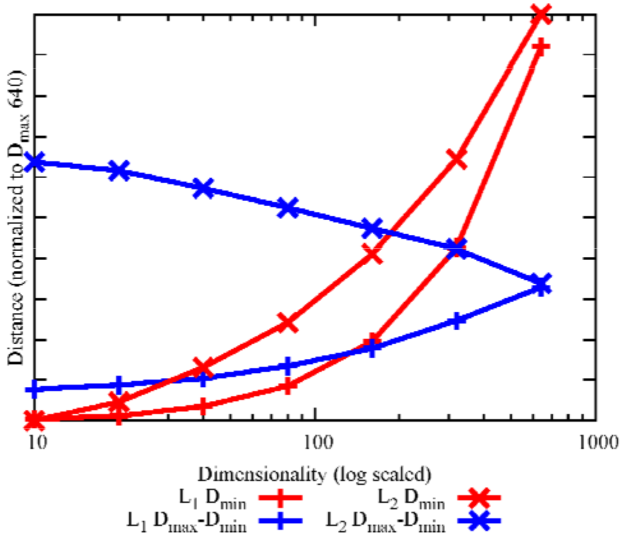


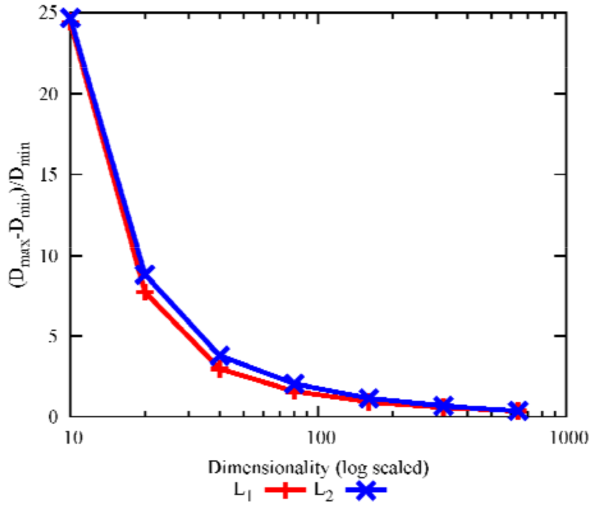
All-Relevant

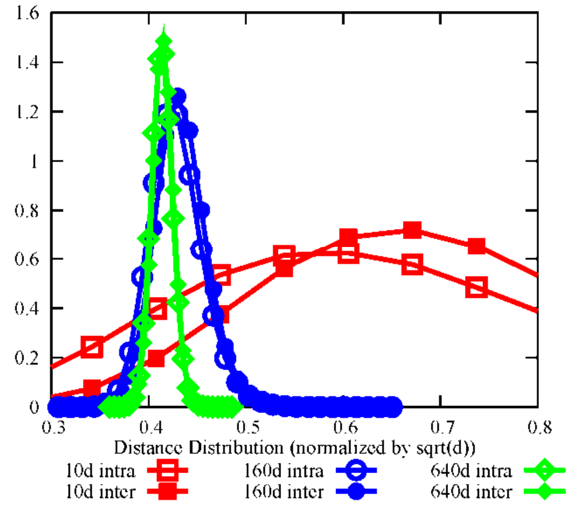
$$\lim_{d\to\infty} \mathrm{var}\left(\frac{\|X_d\|}{E\left[\|X_d\|\right]}\right)=0 \quad\Rightarrow\quad \frac{D_{\max}-D_{\min}}{D_{\min}} \to 0$$

Experimental Set-Up

Data sets show properties of the “curse of dimensionality”





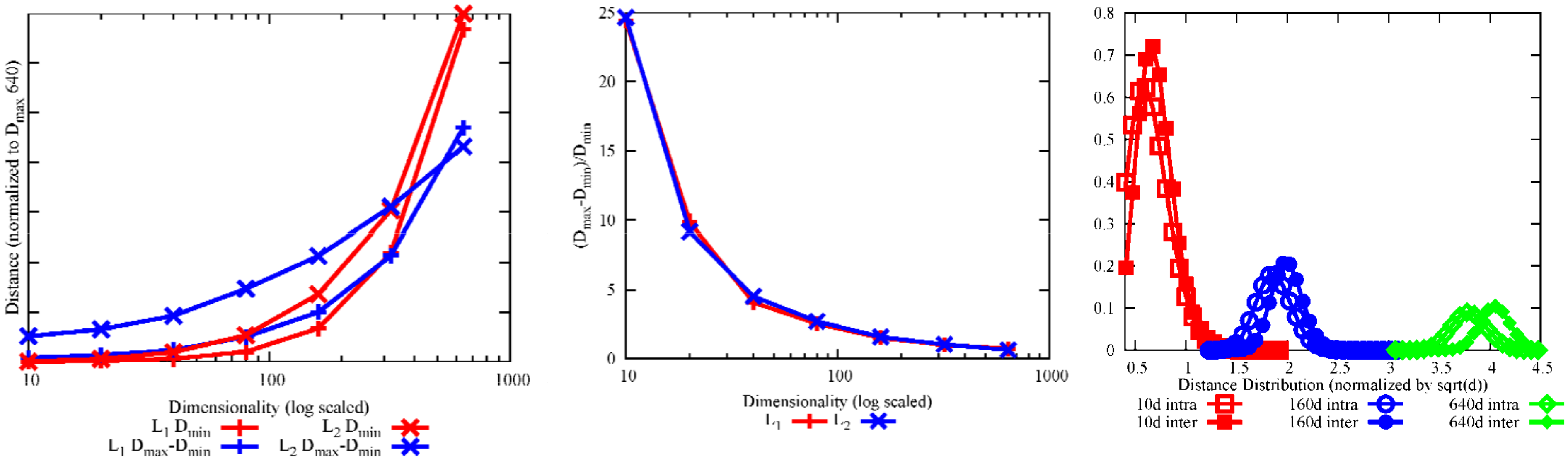


10-Relevant

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

Experimental Set-Up

Data sets show properties of the "curse of dimensionality"

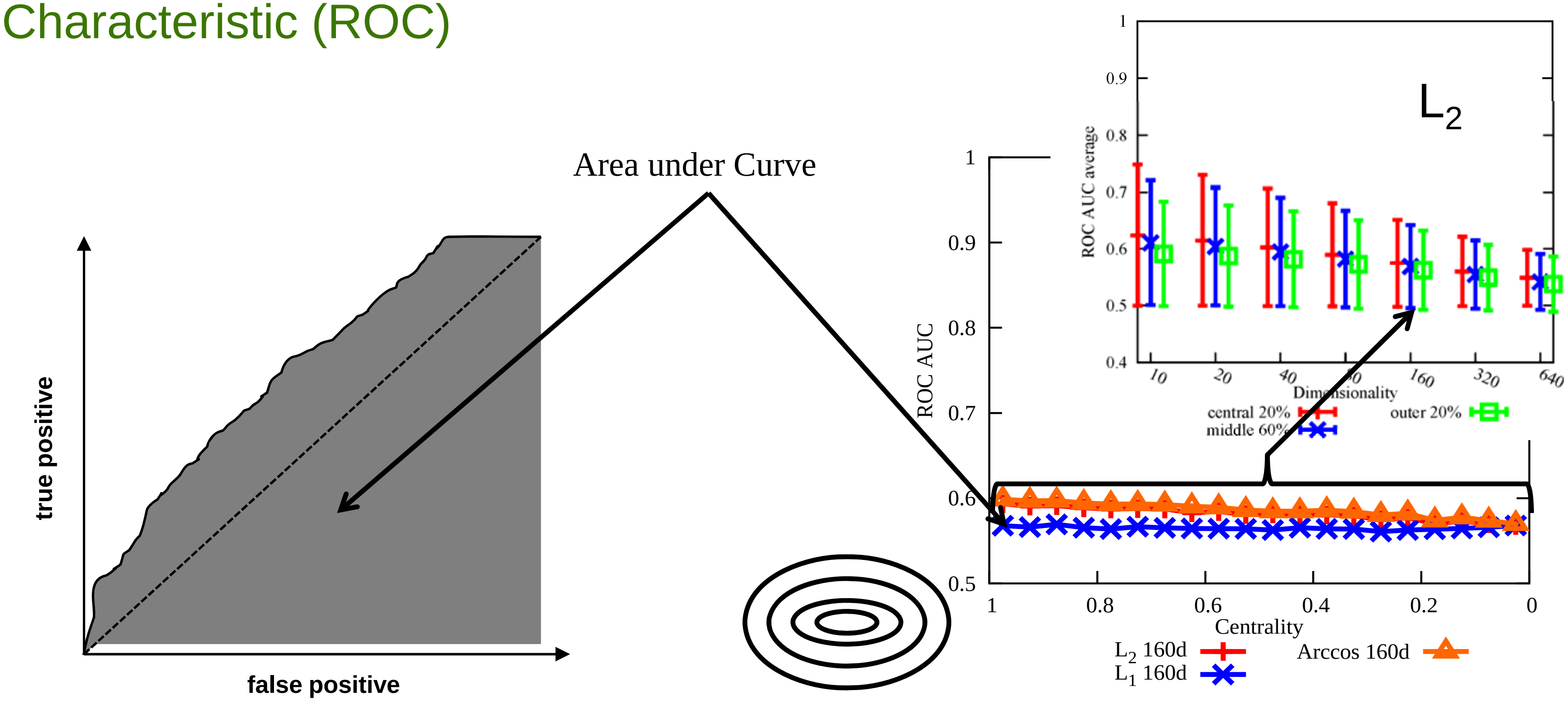


All-Dependent

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

Experimental Set-Up

- 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)

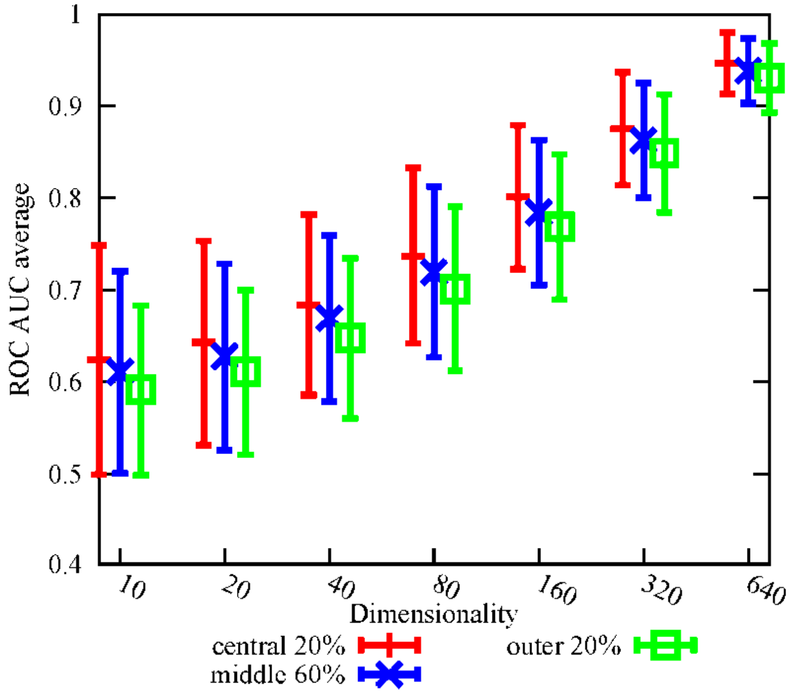


one query (neighbors same/different cluster)

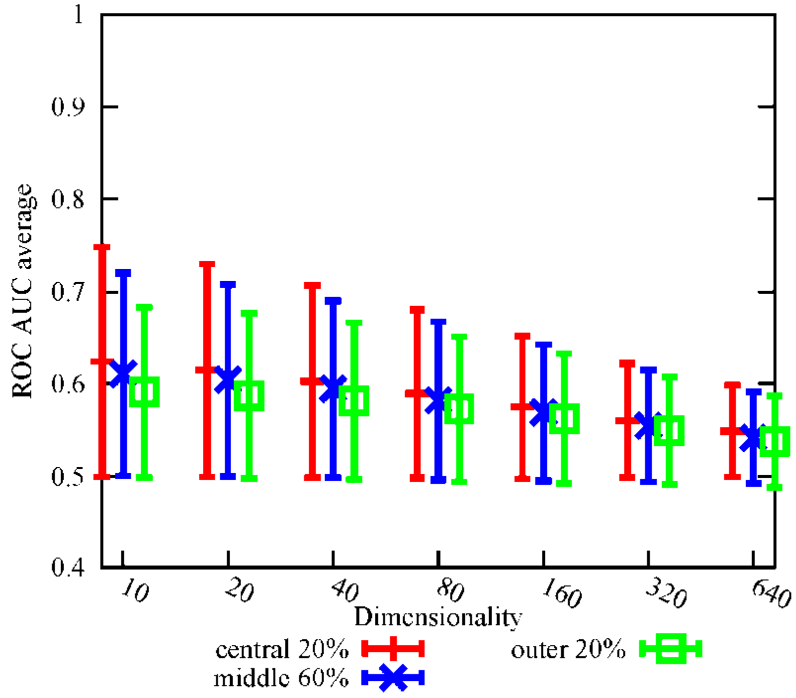
average over all items as a query

Observations

Euclidean distance



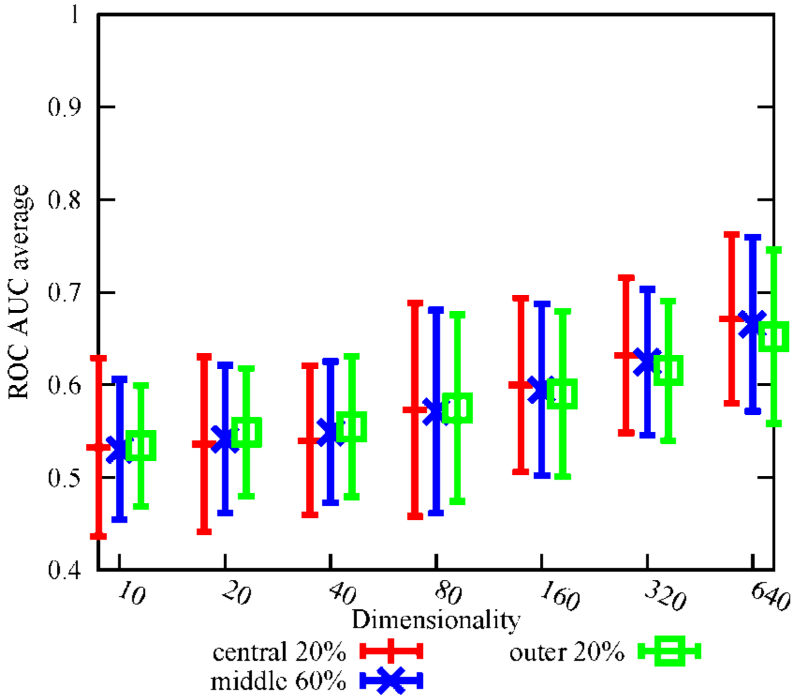
All-Relevant



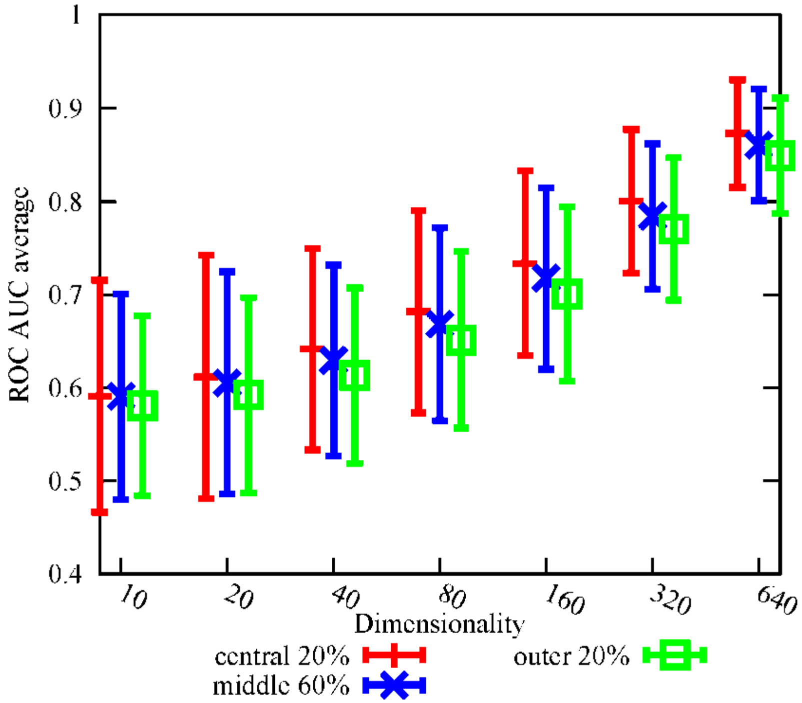
10-Relevant

Observations

Euclidean distance



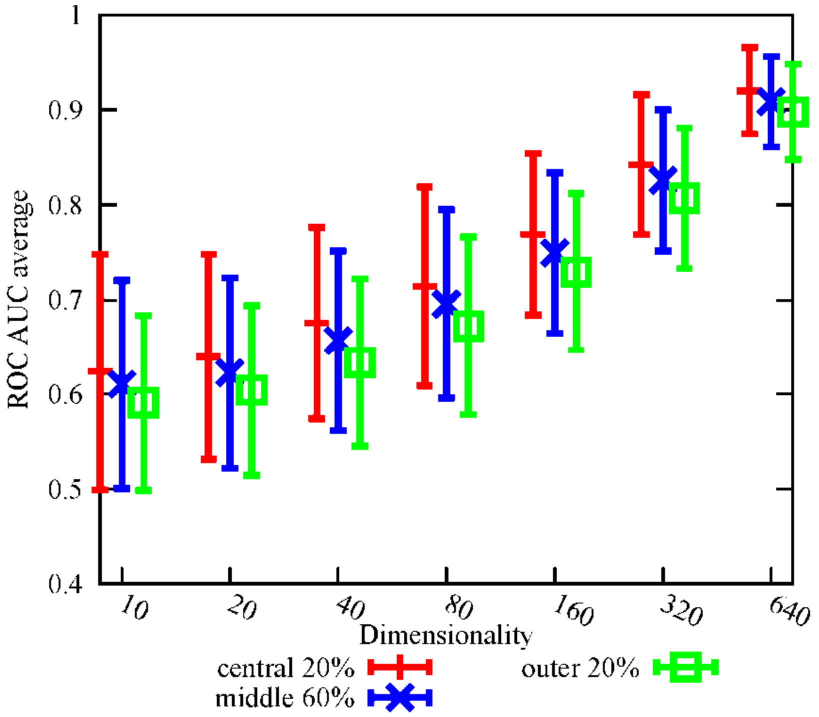
Cyc-Relevant



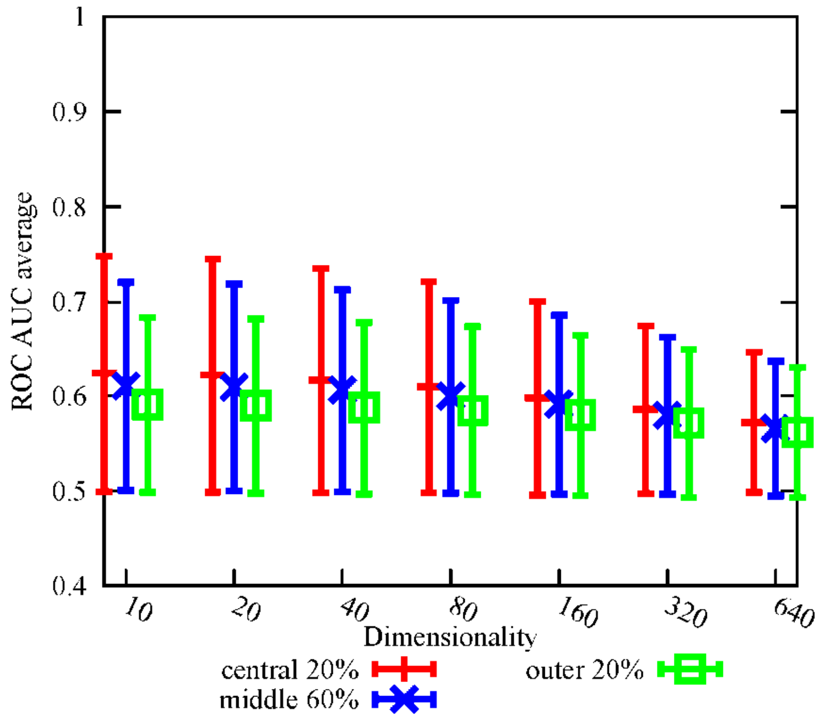
Half-Relevant

Observations

Euclidean distance



All-Dependent

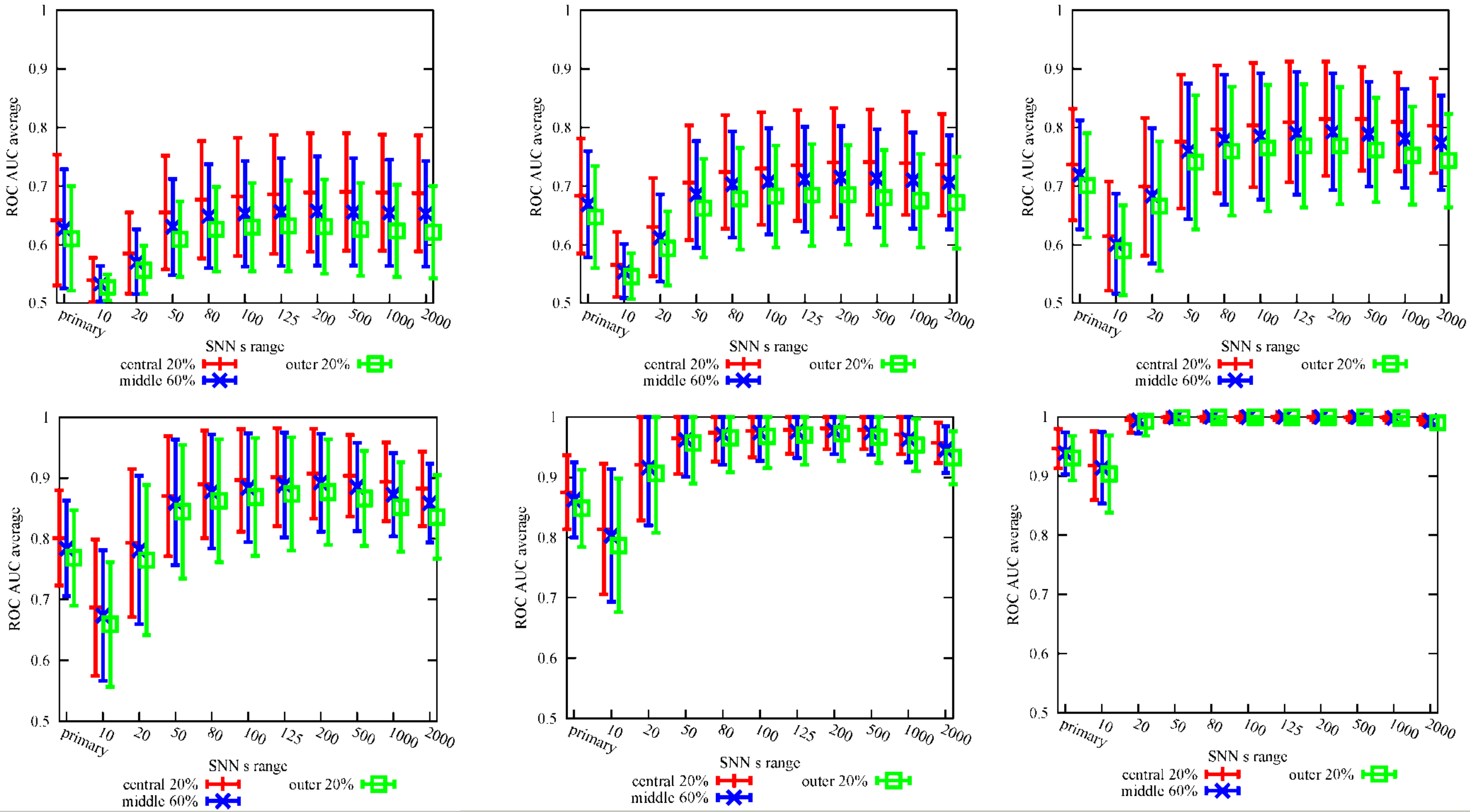


10-Dependent

Observations

SNN based on Euclidean

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

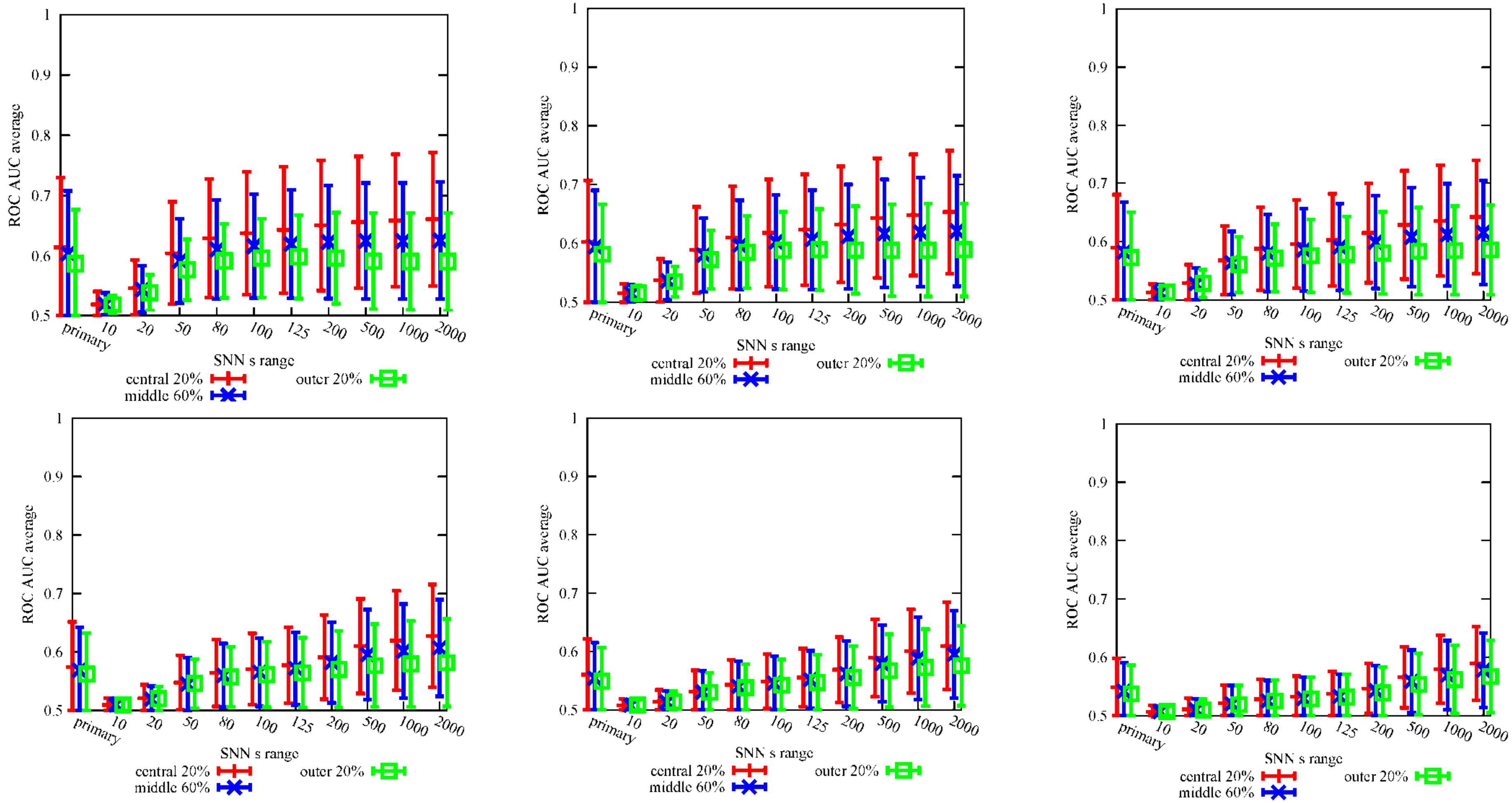


Observations

SNN based on Euclidean

10-Relevant

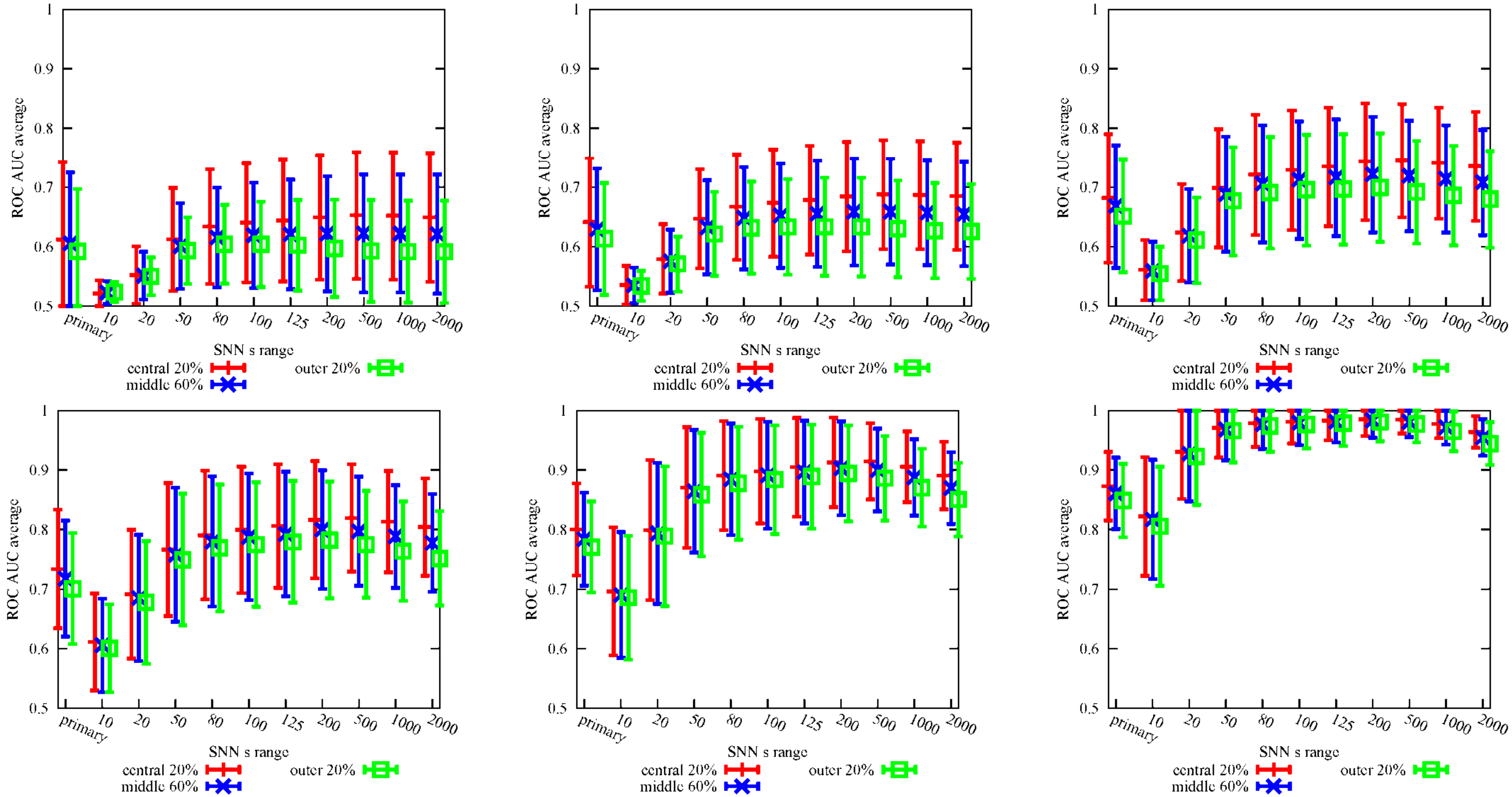
20/40/80/160/320/640 dimensions



Observations

SNN based on Euclidean

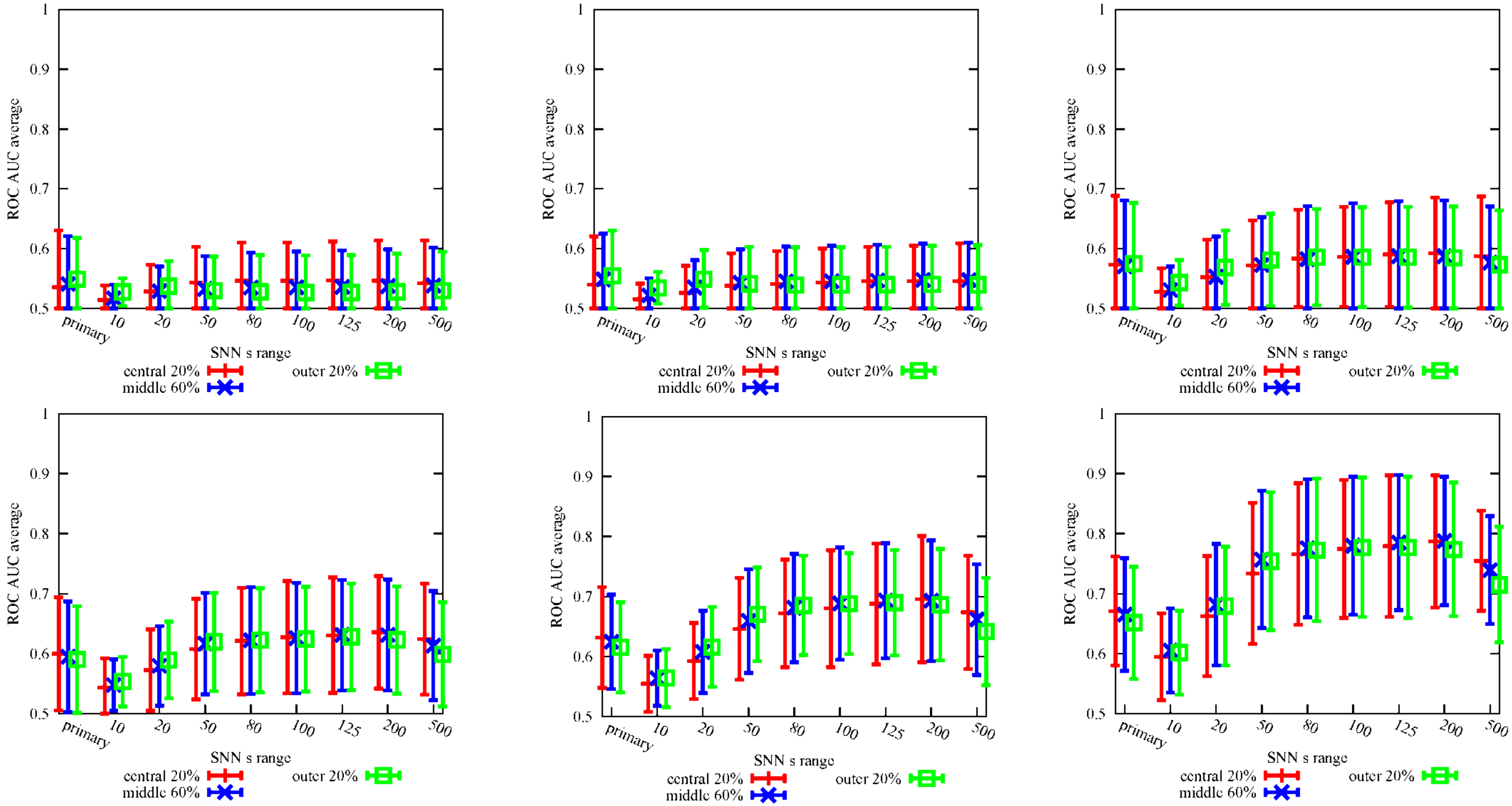
Half-Relevant
20/40/80/160/320/640 dimensions



Observations

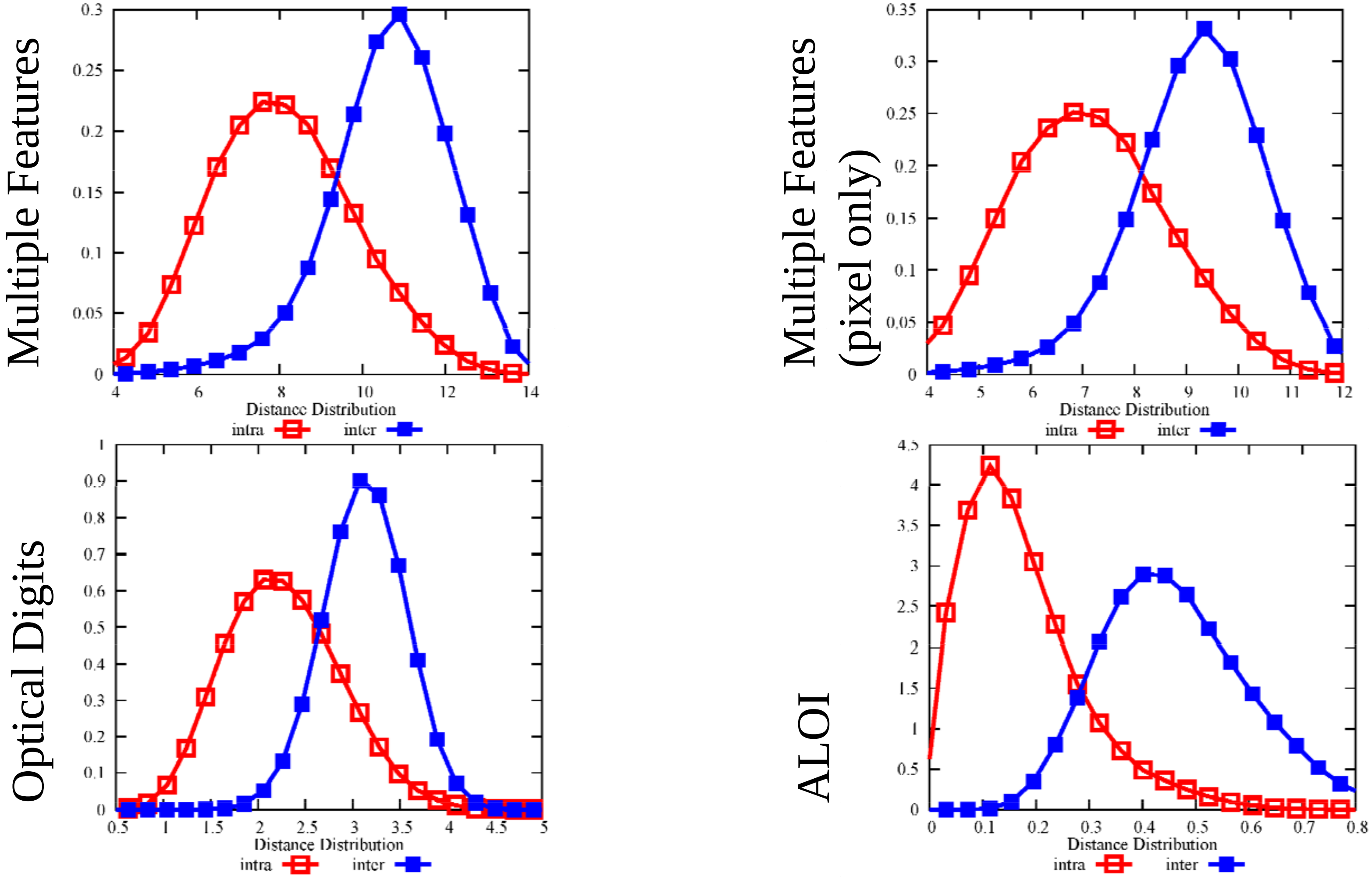
SNN based on Euclidean

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



Observations

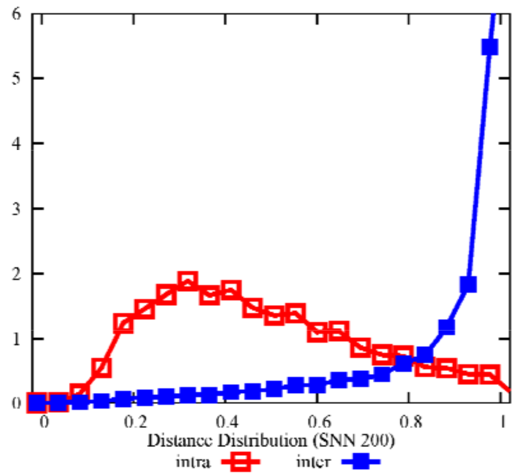
some real data sets: distributions of Euclidean distances



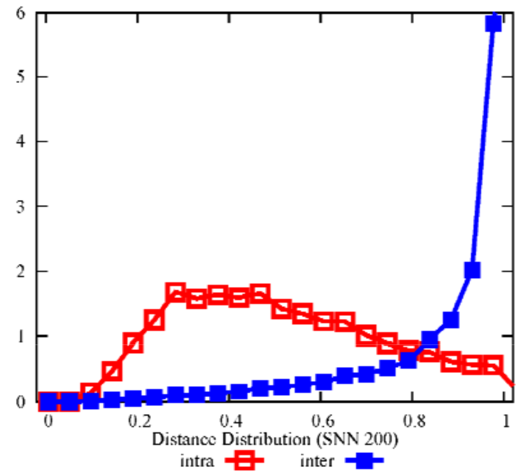
Observations

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

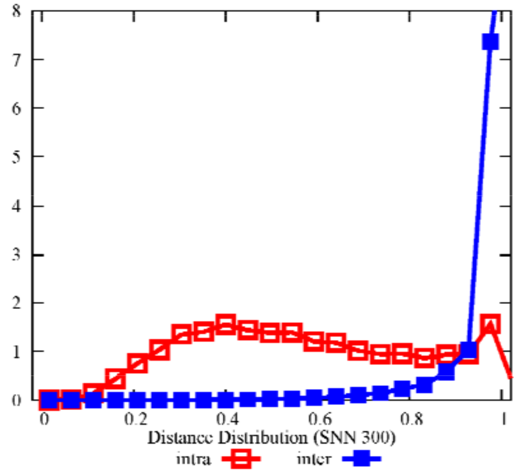
Multiple Features



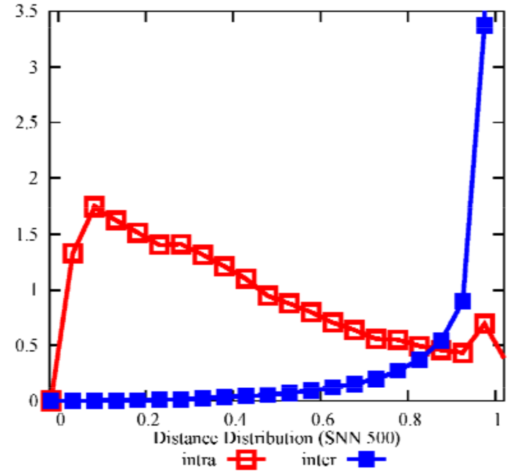
Multiple Features (pixel only)



Optical Digits

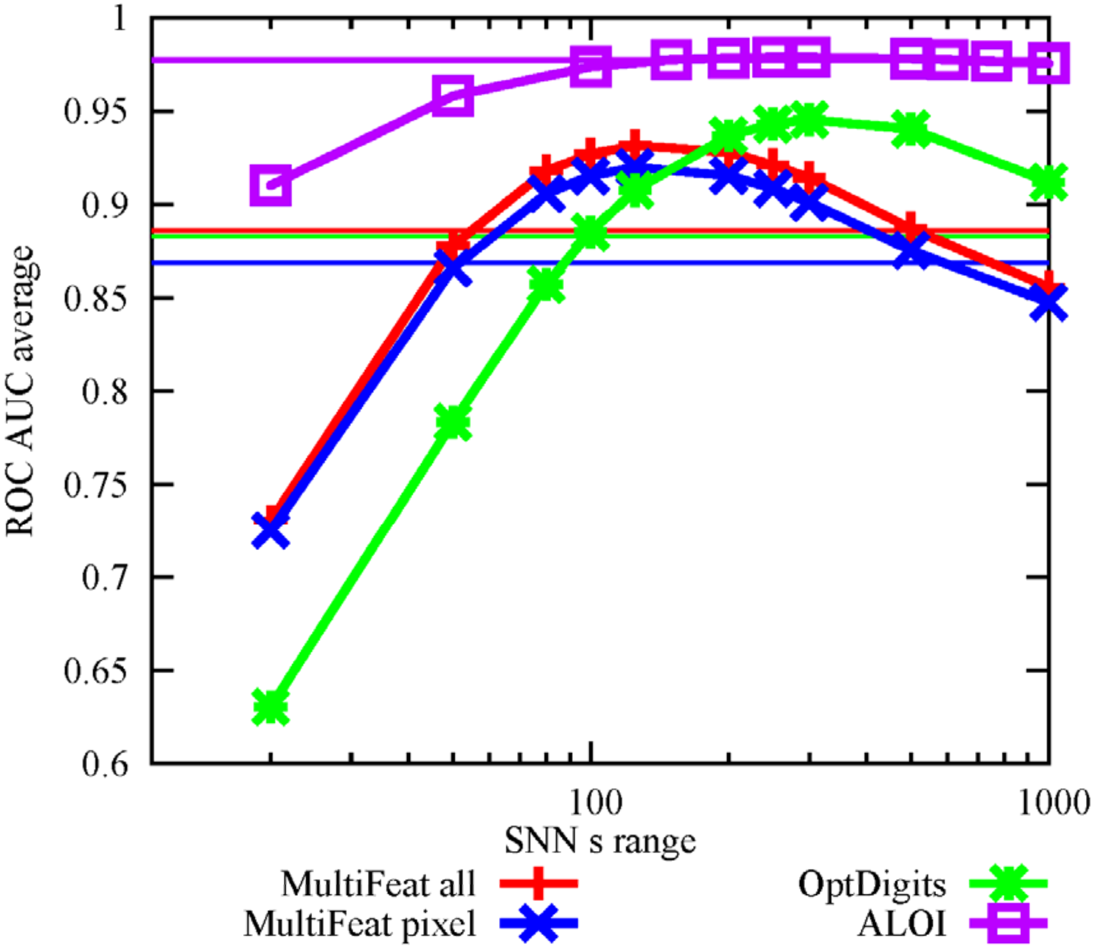


ALOI



Observations

some real data sets: ranking quality



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
- SNN information *can* improve neighborhood ranking for even very low quality of neighborhood queries
  - if the primary distance already performs good, the improvement by SNN in many cases seems actually to be more significant
  - open questions:
    - good choice of neighborhood size *s*: relationship between *s* and size of natural clusters?
    - *k*NN query based on SNN_s: relationship between *k* and *s*?

supplementary material:

http://www.dbs.ifi.lmu.de/research/SNN/