Motivation

Interpreting Distance Fct.
- Distance Functions
- Curse of Dimens.
- SNN Distance

Experiments
- SNN performance
- Histograms
- Effects of noise

Conclusions

Quality of Similarity Rankings in Time Series

12th International Symposium on Spatial and Temporal Databases (SSTD 2011)

Thomas Bernecker[1], Michael E. Houle[2], Hans-Peter Kriegel[1], Peer Kröger[1], Matthias Renz[1], *Erich Schubert*[1], Arthur Zimek[1]

[1] Ludwig-Maximilians-Universität München, Munich, Germany
[2] National Institute of Informatics, Tokyo, Japan

2011-08-26 — Minneapolis, MN

Quality of Similarity Rankings in Time Series

T. Bernecker, M. E. Houle, H.-P. Kriegel, P. Kröger, M. Renz, E. Schubert, A. Zimek

Motivation

Interpreting Distance Fct.
Distance Functions
Curse of Dimens.
SNN Distance

Experiments
SNN performance
Histograms
Effects of noise

Conclusions

Time series research

...has plenty of:

- New distance functions
- Dimensionality reduction
- Approximations

...but:

- How big is a distance of 0.432?
- How big is a difference of 0.123?

What is the meaning of these values?

Interpreting distance functions

T. Bernecker, M. E. Houle, H.-P. Kriegel, P. Kröger, M. Renz, E. Schubert, A. Zimek

Motivation
Interpreting Distance Fct.
Distance Functions
Curse of Dimens.
SNN Distance
Experiments
SNN performance
Histograms
Effects of noise
Conclusions

Distance functions used to have a physical meaning:

- "As the crow flies"
- "Taxicab metric"

This worked well for the three-dimensional world.

But this is not so in time series:

- "Curse of dimensionality"
  loss of contrast in high-dimensional data
- Dimension-alignment as done by time warping
- Edit distances treat big and small edits the same

But: the distance functions work!

The "Curse of Dimensionality"

T. Bernecker, M. E. Houle, H.-P. Kriegel, P. Kröger, M. Renz, E. Schubert, A. Zimek

Motivation
Interpreting Distance Fct.
Distance Functions
Curse of Dimens.
SNN Distance
Experiments
SNN performance
Histograms
Effects of noise
Conclusions

Commonly described as

- Distances become "indiscernible"
- Distances "lose their usefulness"
- Hypercube becomes "vastly" bigger than hypersphere
- Nearest and farthest neighbor become similar
- Mathematical:

$$\lim_{\mathrm{dim}\to\infty} \frac{\mathrm{dist}_{\max} - \mathrm{dist}_{\min}}{\mathrm{dist}_{\min}} \to 0$$

So they *should* not work.
But: they do!

How bad is the "Curse of Dimensionality"?

Some facts on the "Curse of Dimensionality" (from Houle et al. 2010):

- Mathematics proven for i.i.d. data only
- Relevant dimensions make the problem *easier*
- Irrelevant dimensions make the problem harder
- $\Rightarrow$ mostly a matter of "signal to noise ratio"
- Numerical contrast goes away,
  but *ranking* still remains meaningful

Goal: Restore contrast and intuition using the ranking information of the existing distance functions!

Shared Nearest Neighbor Similarity

Idea: Similar objects have similar neighbors.

$$\begin{aligned} SNN_s(x,y) &= |NN_s(x) \cap NN_s(y)| \\ simcos_s(x,y) &= \frac{SNN_s(x,y)}{s} \end{aligned}$$

Properties:

- Intuitive value range from “None” to “All”
- Intuitive interpretation (“social”)
- Good contrast, good performance
- Needs an “okay” existing ranking
- Extra parameter s to choose
- More expensive to use (second order distance)

Shared Nearest Neighbor Distance

T. Bernecker, M. E. Houle, H.-P. Kriegel, P. Kröger, M. Renz, E. Schubert, A. Zimek

Motivation
Interpreting Distance Fct.
Distance Functions
Curse of Dimens.
SNN Distance
Experiments
SNN performance
Histograms
Effects of noise
Conclusions

The similarity function needs to be transformed to a (non-metrical) distance function:

$$\begin{aligned} dinv_s(x,y) &= 1 - simcos_s(x,y) \\ dacos_s(x,y) &= \arccos\left(simcos_s(x,y)\right) \\ dln_s(x,y) &= -\ln simcos_s(x,y) \end{aligned}$$

Just like cosine distance.
Interpretable as “cosine distance” in “neighbor space”.
Similar: Jaccard distance (metrical)

$$J(x,y) := 1 - \frac{|NN_s(x) \cap NN_s(y)|}{|NN_s(x) \cup NN_s(y)|}$$

Quality of Similarity Rankings in Time Series

T. Bernecker, M. E. Houle, H.-P. Kriegel, P. Kröger, M. Renz, E. Schubert, A. Zimek

Motivation

Interpreting Distance Fct.
- Distance Functions
- Curse of Dimens.
- SNN Distance

Experiments
- SNN performance
- Histograms
- Effects of noise

Conclusions

Experimental results

Data sets used

Four very different data sets:

- Cylinder-Bell-Funnel (CBF): artificial
- Synthetic control: artificial
- Leaf dataset: outlines of tree leafs
- Lightning-7: lightning strike emissions

Each modified in different ways:

- Original data set
- Extended with noise (irrelevant attributes)
- Extended with “signal” (relevant attributes)

Quality of Similarity Rankings in Time Series

T. Bernecker, M. E. Houle, H.-P. Kriegel, P. Kröger, M. Renz, E. Schubert, A. Zimek

Motivation

Interpreting Distance Fct.
Distance Functions
Curse of Dimens.
SNN Distance

Experiments
SNN performance
Histograms
Effects of noise

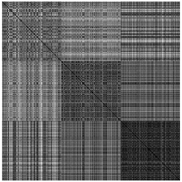
Conclusions

Results on unmodified data sets

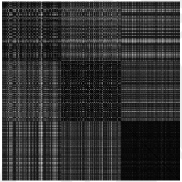
Benefits of using SNN
Exemplary on the Cylinder-Bell-Funnel (artificial) data set

Contrast gain using SNN

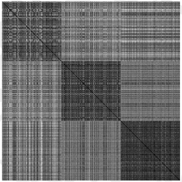
Visual improvement (unmodified CBF data set):



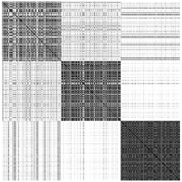
Euclidean



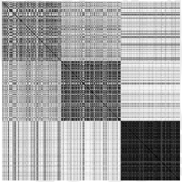
DTW 20%



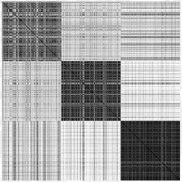
LCSS 20%



DTW $s = 70$



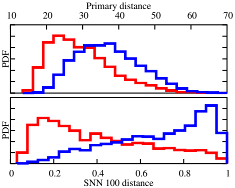
DTW $s = 100$



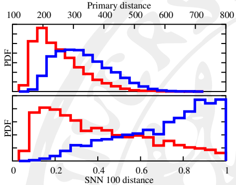
LCSS $s = 100$

Distance Histograms

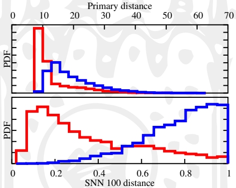
Numerical contrast improved (unmodified CBF data set):



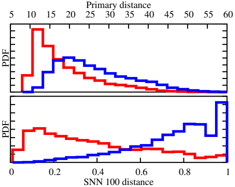
Euclidean



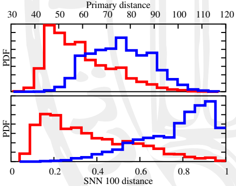
Manhattan



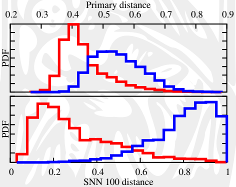
DTW 20%



ERP 20%



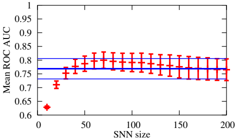
EDR 20%



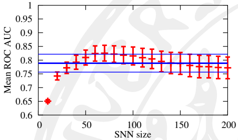
LCSS 20%

Effect of neighborhood size s:

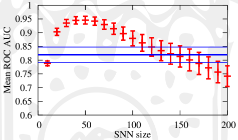
Effect of variation of SNN size parameter s (CBF):



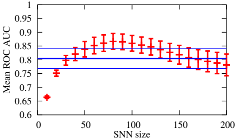
Euclidean



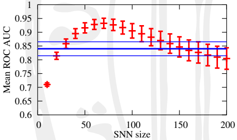
Manhattan



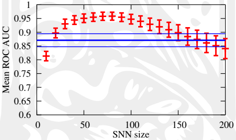
DTW 20%



ERP 20%



EDR 20%



LCSS 20%

Quality of Similarity Rankings in Time Series

T. Bernecker, M. E. Houle, H.-P. Kriegel, P. Kröger, M. Renz, E. Schubert, A. Zimek

Motivation

Interpreting Distance Fct.
- Distance Functions
- Curse of Dimens.
- SNN Distance

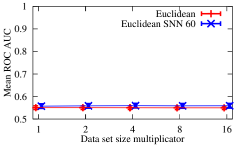
Experiments
- SNN performance
- Histograms
- Effects of noise

Conclusions

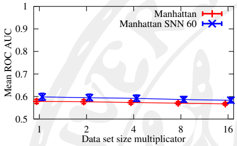
Results on modified data sets

Adding noise to the data set,
Changing the signal to noise ratio

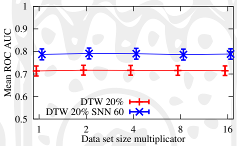
Adding noise to the data (Leaf data set)



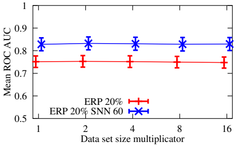
Euclidean



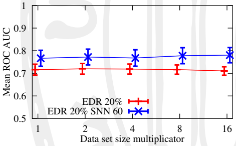
Manhattan



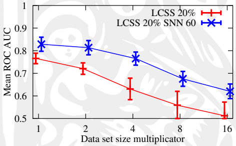
DTW 20%



ERP 20%



EDR 20%

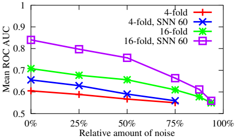


LCSS 20%

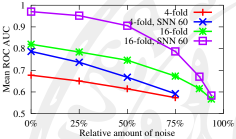
T. Bernecker, M. E. Houle, H.-P. Kriegel, P. Kröger, M. Renz, E. Schubert, A. Zimek

Motivation
Interpreting Distance Fct.
Distance Functions
Curse of Dimens.
SNN Distance
Experiments
SNN performance
Histograms
Effects of noise
Conclusions

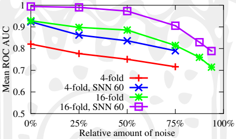
Changing the signal-to-noise ratio (Leaf data set)



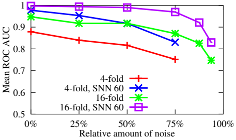
Euclidean



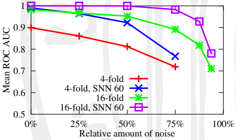
Manhattan



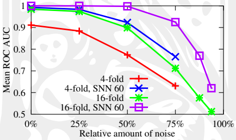
DTW 20%



ERP 20%



EDR 20%



LCSS 20%

Conclusions

Second order "shared nearest neighbor" distances offer:

- Improved performance
- Better numerical contrast
- Parameter s is not difficult to choose
- Less sensitive to noise
- ... but computationally more expensive

Motivation

Interpreting Distance Fct.
- Distance Functions
- Curse of Dimens.
- SNN Distance

Experiments
- SNN performance
- Histograms
- Effects of noise

Conclusions

Thank you for your attention!