Outlier Detection in Axis-Parallel Subspaces of High Dimensional Data

PAKDD 2009

Hans-Peter Kriegel, Peer Kröger, Erich Schubert, Arthur Zimek

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

{kriegel,kroegerp,schube,zimek}@dbs.ifi.lmu.de

Outline

1. Motivation
2. Subspace Outlier
3. Reference Set for Outliers
4. Comparison to Existing Approaches
5. Conclusion

Outline

1. Motivation
2. Subspace Outlier
3. Reference Set for Outliers
4. Comparison to Existing Approaches
5. Conclusion

Motivation

- Hawkins Definition:

  "An outlier is an observation which deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism."

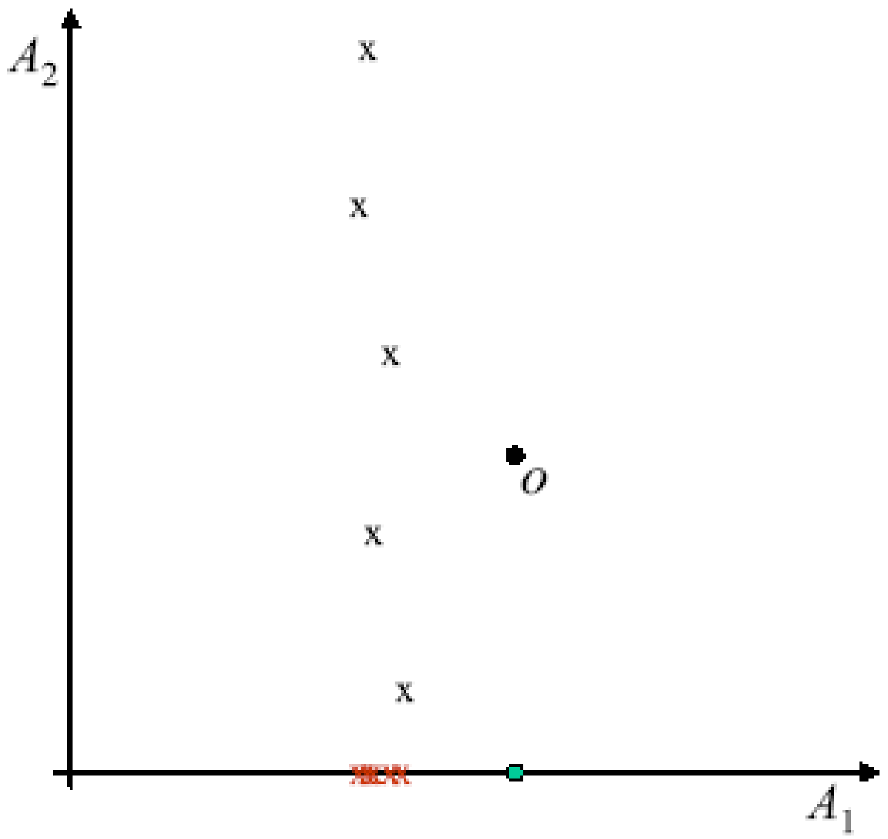
- Collecting data with high dimensionality

  → "*curse of dimensionality*"

- two aspects here:
  - Euclidean distances (as commonly used) loose their expressiveness: no outlier can be detected that deviates considerably from the majority of points in comparison to other points
  - a "generating mechanism" to identify may be responsible for a subset of the features only (*local feature relevance*)

Motivation

- try to find outliers in subspaces, i.e., based on the subset of features related to a "generating mechanism"
- subspace $\{A_1\}$:
  o is an outlier
- subspace $\{A_2\}$:
  o is not an outlier
- full-dimensional space $\{A_1, A_2\}$:
  o is not an outlier



- distribution of attribute values in A_2 appears to be not relevant for the "mechanism" in question

Outline

1. Motivation
2. Subspace Outlier
3. Reference Set for Outliers
4. Comparison to Existing Approaches
5. Conclusion

Subspace Outlier

general idea:

- assign a set of reference points to a point o

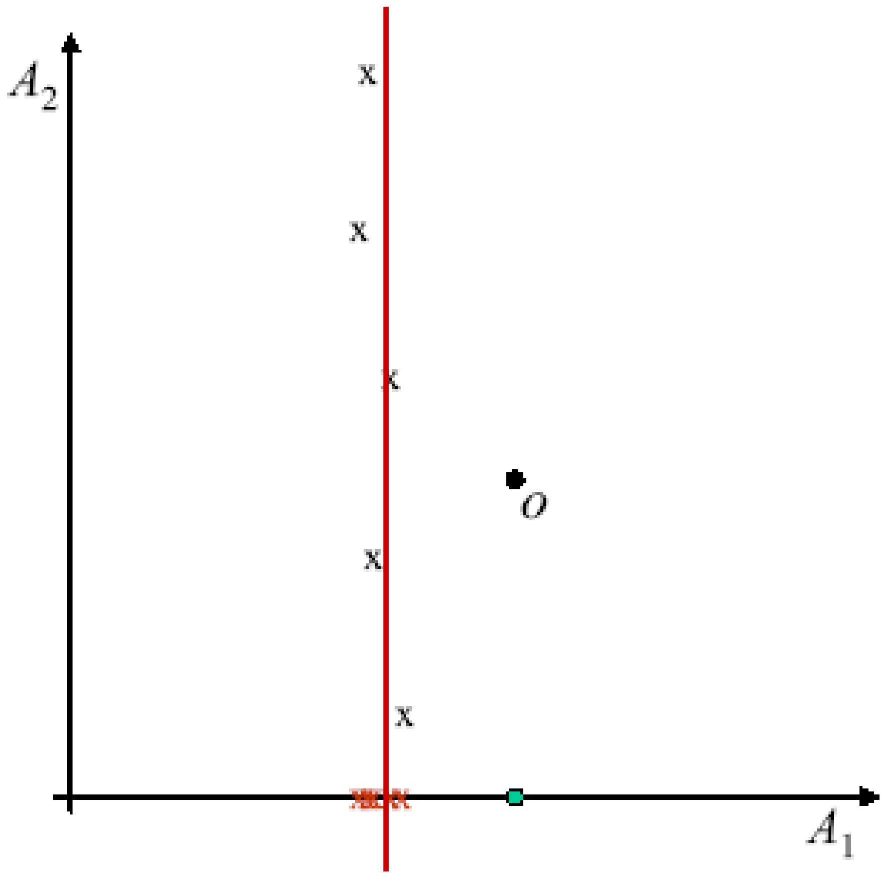
  (e.g., k-nearest neighbors – but keep in mind the “curse of dimensionality”: local feature relevance vs. meaningful distances)

- find the subspace spanned by these reference points (allowing some jitter)

- analyze for the point o how well it fits to this subspace

Subspace Outlier

- subspace spanned a set of points S: orthogonal to a subspace minimizing the variance but maximizing the number of attributes - a hyperplane more or less accommodating the set S of reference points
- within this subspace, the variance of the points in S is high
- in the perpendicular space, the variance is low



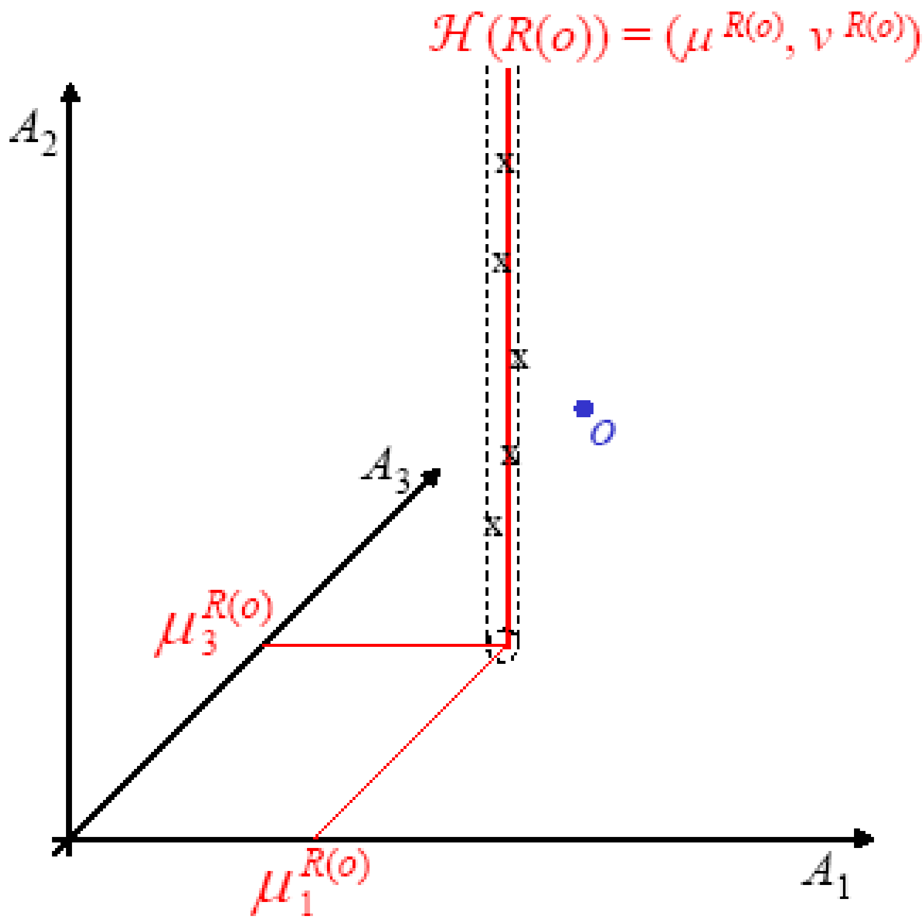
Subspace Outlier

- variance VAR^S: averaged square distance of the points in S to the mean μ^S:

$$VAR^S = \frac{\sum_{p \in S} \left(dist\left(p, \mu^S\right)\right)^2}{|S|}$$

- variance along attribute i:

$$\text{var}_i^S = \frac{\sum_{p \in S} \left(dist\left(p_i, \mu_i^S\right)\right)^2}{|S|}$$



Subspace Outlier

- derive the subspace: subspace defining vector specifies the relevant attributes of the subspace defined by a reference set, i.e., the attributes where the reference points exhibit low variance
- in all *d* attributes, the points have a total variance of VAR^S
- the expected variance along attribute i is VAR^S / d
- variance along attribute *i* is *low* if it is smaller than the expected variance by a predefined coefficient α:

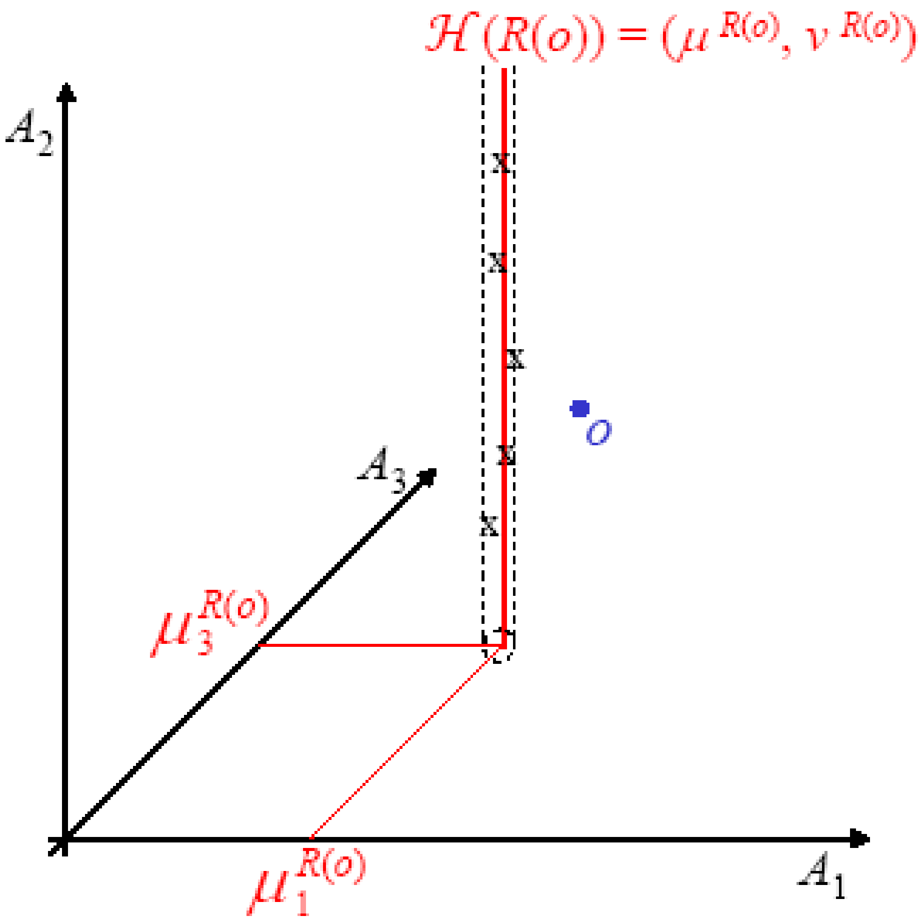
$$v_i^S = \begin{cases} 1, & if\ \mathrm{var}_i^S < \alpha \dfrac{VAR^S}{d} \\ 0, & else \end{cases}$$

Subspace Outlier

- subspace hyperplane *H*(S) of reference set S is defined by mean value μ^S and the subspace defining vector v^S

- points in the reference set *R(o)* of *o* form a line in three-dimensional space

  $v^{R(o)} = (1,0,1)$

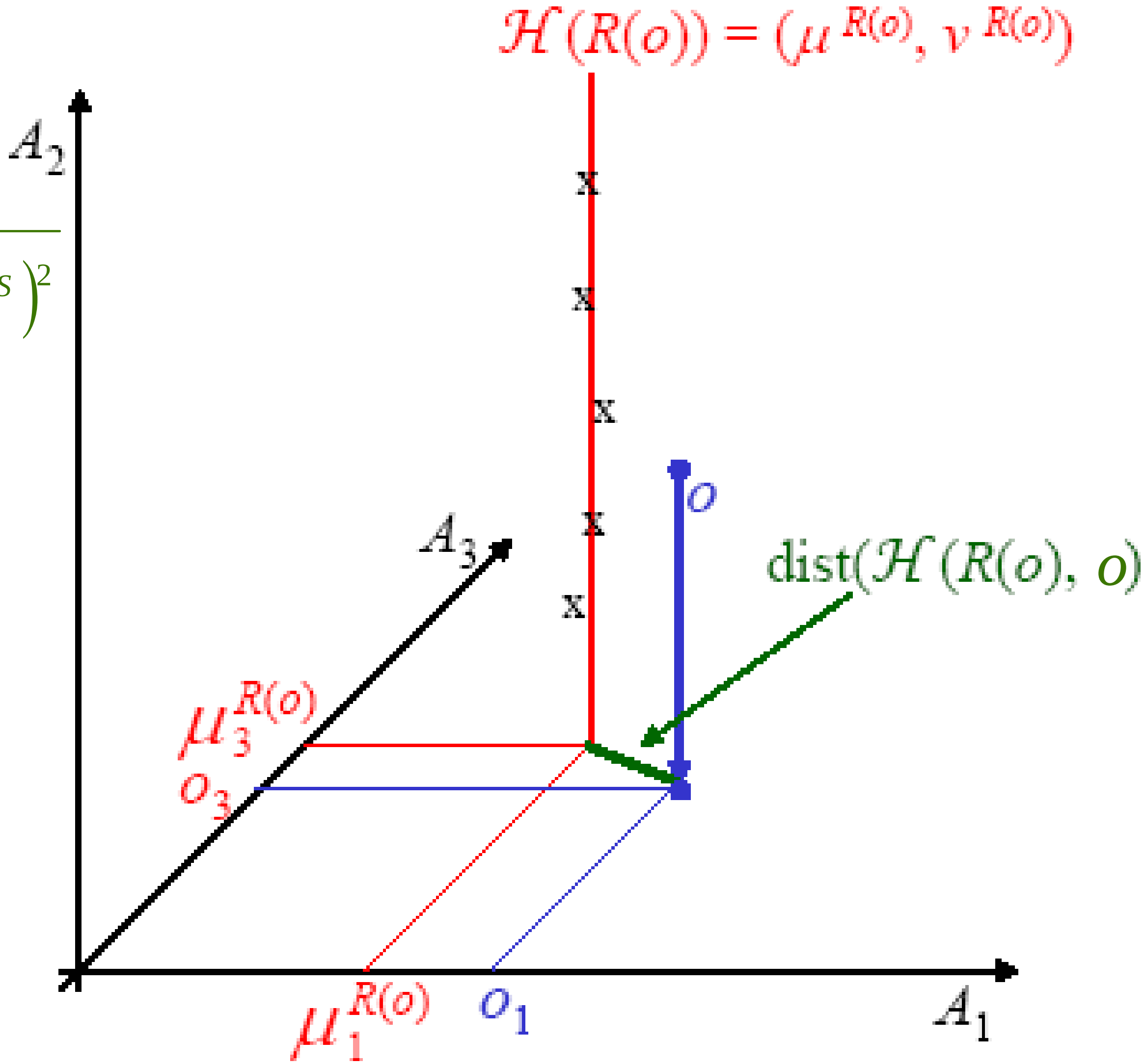


Subspace Outlier

- distance of *o* to the reference hyperplane:

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

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



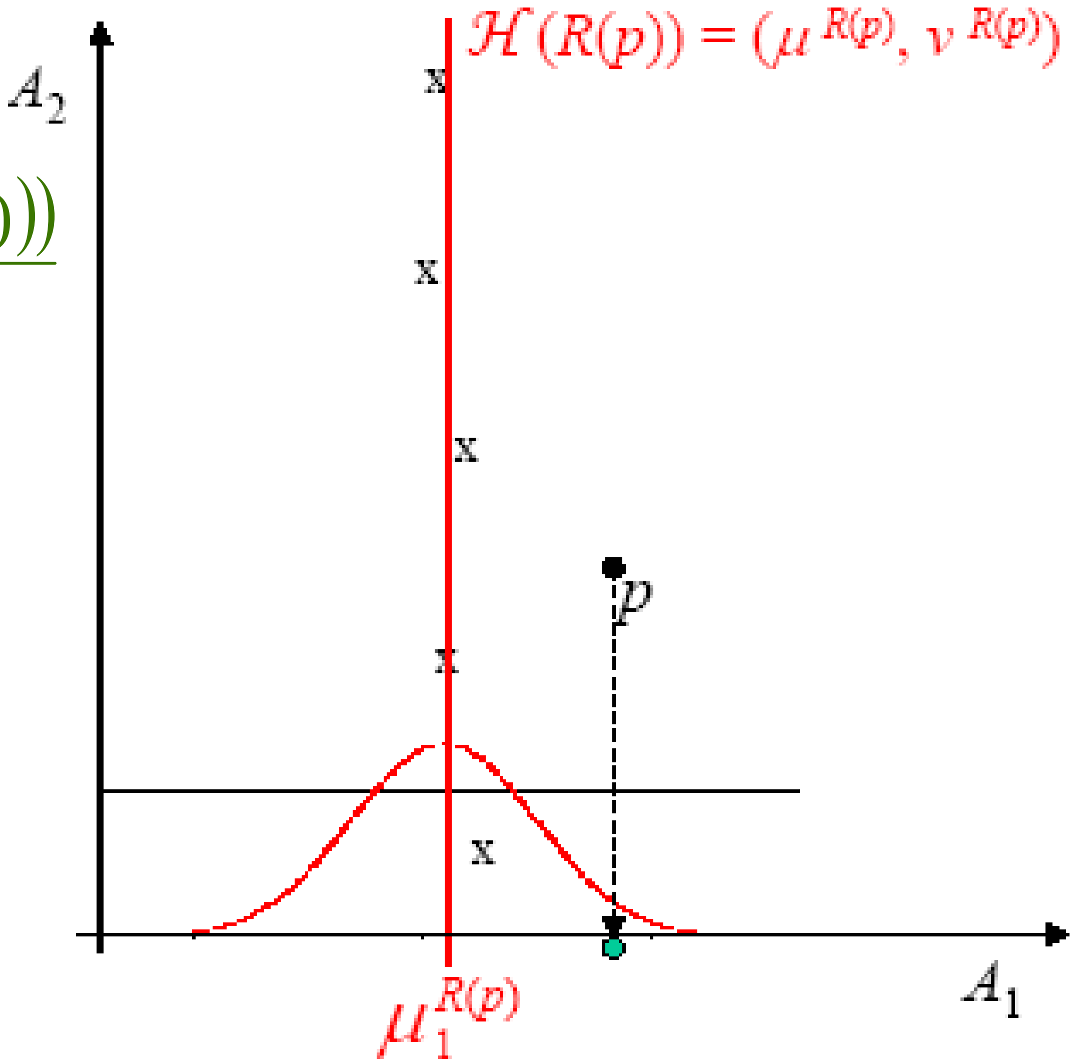
Subspace Outlier

subspace outlier degree (SOD) of a point p:

$$SOD_{R(p)}(p) = \frac{dist\big(p, H\big(R(p)\big)\big)}{\left|v^{R(p)}\right|}$$

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



Outline

1. Motivation
2. Subspace Outlier
3. Reference Set for Outliers
4. Comparison to Existing Approaches
5. Conclusion

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 *l* nearest neighbors in terms of the shared nearest neighbor similarity
  - given a primary distance function *dist* (e.g. Euclidean distance)
  - $N_k(p)$: *k*-nearest neighbors in terms of *dist*
  - SNN similarity for two points *p* and *q*:

$$sim_{SNN}(p,q) = \left| N_k(p) \cap N_k(q) \right|$$

  - reference set *R(p)*: *l*-nearest neighbors of *p* using sim_{SNN}
- observations back the assumption that SNN stabilizes neighborhood in high dimensional data

Outline

1. Motivation
2. Subspace Outlier
3. Reference Set for Outliers
4. Comparison to Existing Approaches
5. Conclusion

Comparison to Existing Approaches

complexity:

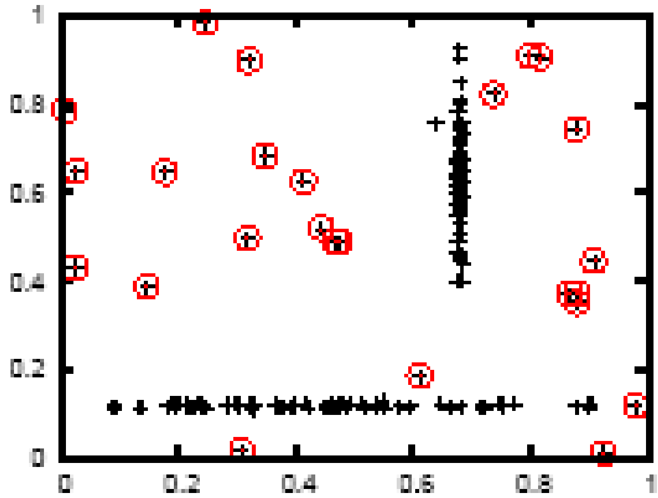
- determine set of *k*-nearest neighbors for each of *n* points:
  $O(dn^2)$
- determine reference set for each point
  (*l*-nearest neighbors based on sim_{SNN}):
  $O(kn)$
- overall (since $k \ll n$):
  $O(dn^2)$

→ comparable to most existing outlier detection algorithms

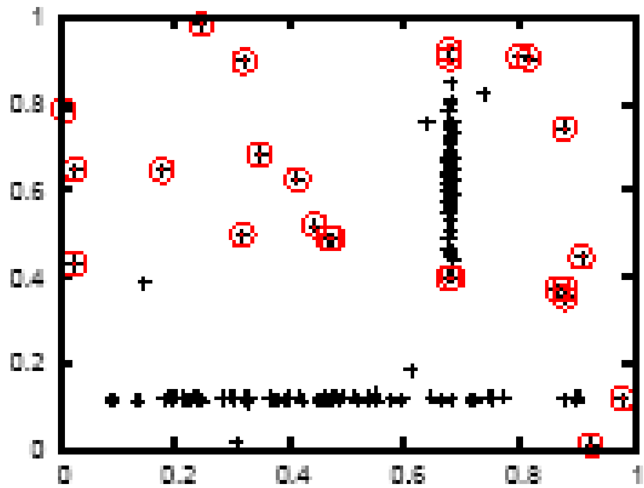
Comparison to Existing Approaches

- 2-d sample data:

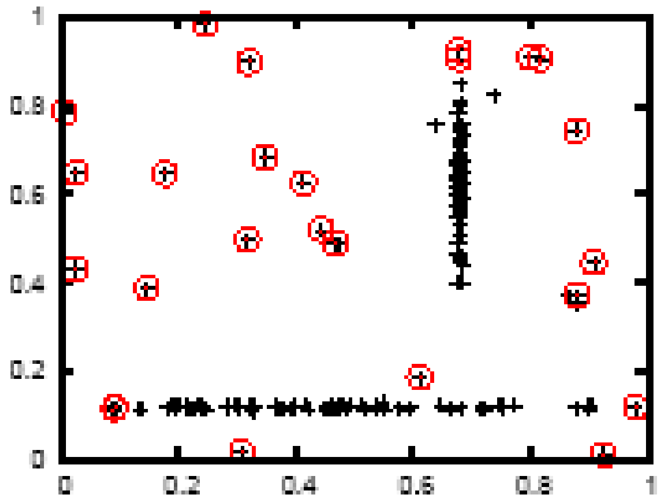
SOD



LOF

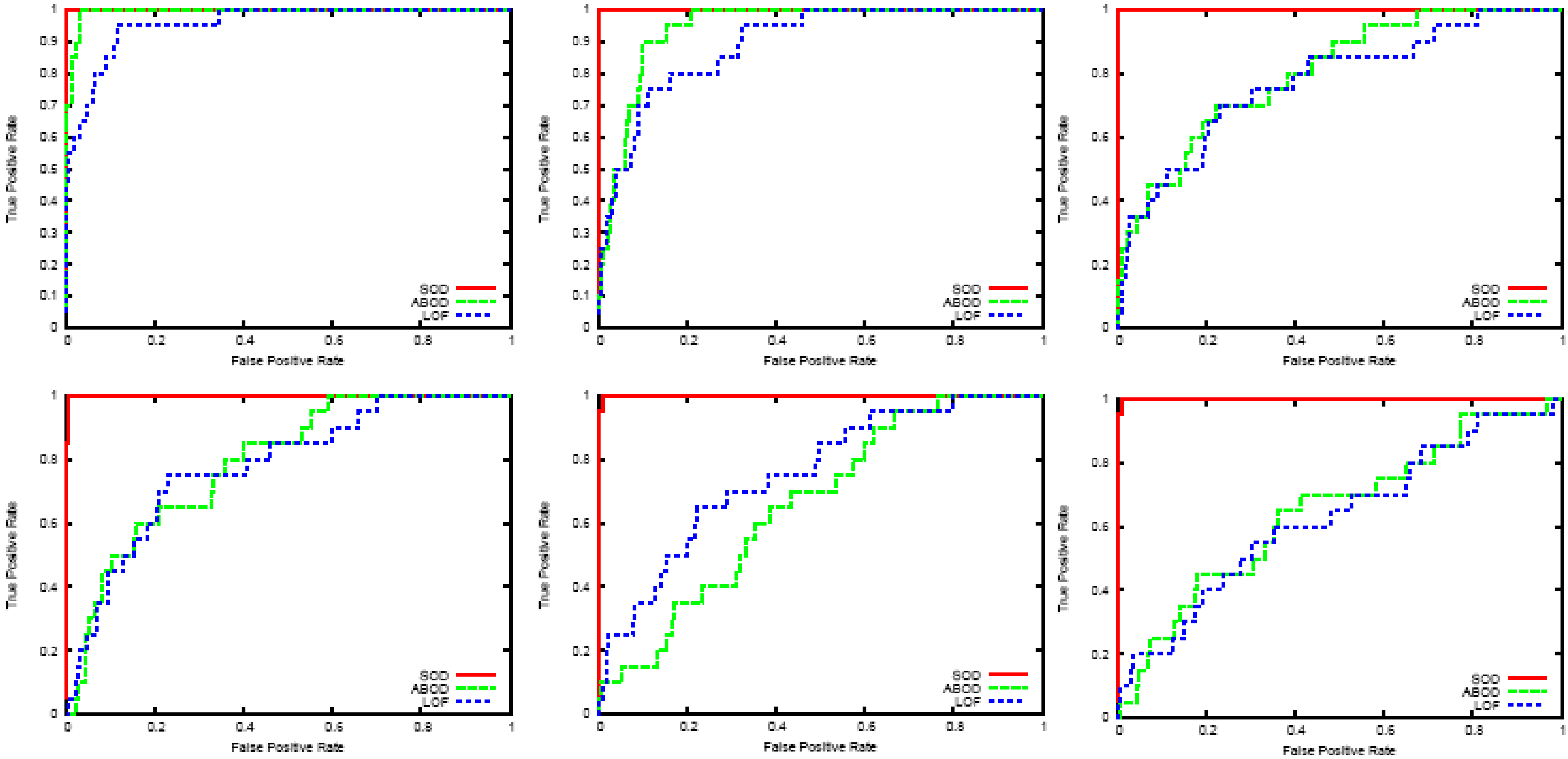


ABOD



Comparison to Existing Approaches

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



Outline

1. Motivation
2. Subspace Outlier
3. Reference Set for Outliers
4. Comparison to Existing Approaches
5. Conclusion

Conclusion

- SOD is a new approach to model outliers in high dimensional data.
- SOD explores outliers in subspaces of the original feature space by combining the tasks of outlier detection and finding the relevant subspace.
- SOD is relatively stable with increasing dimensionality by determining the set of locally relevant neighbors based on SNN.
- SOD finds interesting and meaningful outliers in high dimensional data based on a different intuition compared to full-dimensional outlier models — without adding computational costs.