Universiteit Antwerpen

Karlsruhe Institute of Technology

Flexible Subspace Search for Outlier Detection and Description

Emmanuel Müller

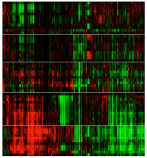
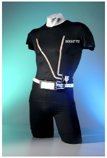
University of Antwerp, Belgium

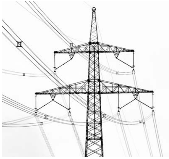
Karlsruhe Institute of Technology, Germany

Dutch-Belgian Database Day (DBDBD 2013), Rotterdam
November 29th, 2013

Outlier Mining Examples

Today's applications provide large and high dimensional databases...



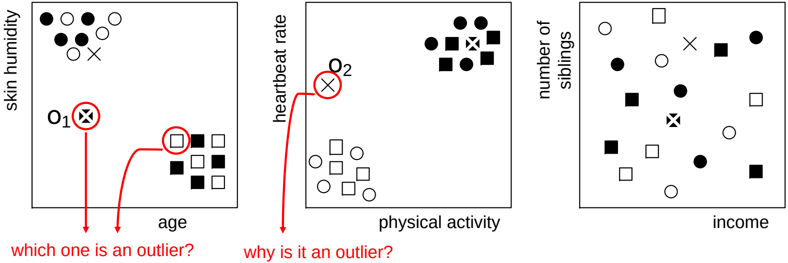
Challenging Databases (e.g. sensor networks)

- Millions of objects, thousands of **attributes per object**
- More and more attributes are measured and stored
- **Loss of contrast**: all objects become unique

⇒ Traditional techniques are insufficient for high dimensional data

Our Solution – Subspaces

Outlier mining in subsets of the given attributes



Subspaces: Relevant Attribute Combinations

- High contrast between outliers and clusters (enable detection)
- Indicate the reasons for high deviation (enable descriptions)

⇒ How to detect such high contrast subspaces?

Overview

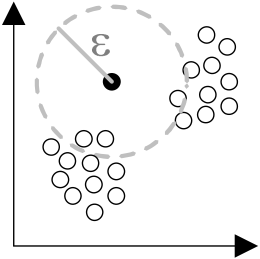
1. Problem Setting: Subspace Search
2. Flexible Subspace Search (RefOut)
3. Evaluation, Application, and Extension
4. Conclusion and Outlook

One of the Traditional Outlier Definitions

- Based on (dis-)similarity of objects w.r.t. **all given dimensions**
- Measure deviation of outlier w.r.t. **all given dimensions**

Density-Based Outliers

- **Underlying density definition**
  $den(o) = |\{p \mid dist(o,p) \leq \varepsilon\}|$
- Outliers have low density in contrast to their densely clustered neighborhood



Outlier Ranking (e.g. Local Outlier Factor[1])

- Sorted list of objects according to **local degree of deviation**
  $\forall o \in DB \;:\; score(o) = 0 \ldots 1$

[1] Breunig, Kriegel, Ng, Sander: **LOF: Identifying density-based Local Outliers**, in ACM SIGMOD, 2000

Subspace Outlier Definitions

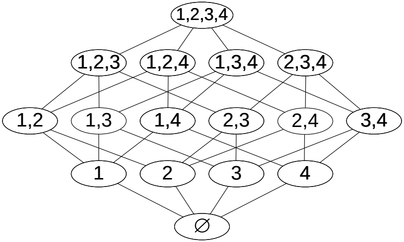
Simple Subspace Definitions

- Utilize traditional definitions in **random subspaces**[2]

$$RS(o) \subseteq \mathcal{P}(D)$$

- Aggregate scores:

$$r(o) = \prod_{S \in RS(o)} score(o, S)$$



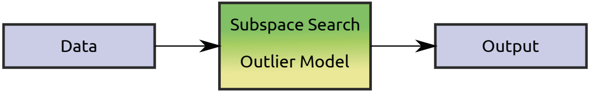
$$dist_S(o,p) = \sqrt{\sum_{i \in S}(o_i - p_i)^2}$$

- Enhanced Subspace Outlier Mining Techniques[3][4][5]

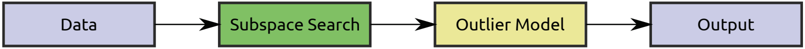
[2] Lazarevic and Kumar: **Feature bagging for outlier detection**, in ACM SIGKDD, 2005.
[3] Müller, Schiffer and Seidl: **Statistical Selection of Relevant Subspace Projections for Outlier Ranking**, in IEEE ICDE 2011.
[4] Keller, Müller and Böhm: **HiCS: High Contrast Subspaces for Density-Based Outlier Ranking**, in IEEE ICDE 2012.
[5] Keller, Müller, Wixler and Böhm: **Flexible and Adaptive Subspace Search for Outlier Analysis**, in ACM CIKM 2013.

Related Work

- Subspace Outlier Mining (e.g. OutRes[3])

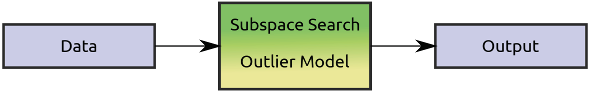


- Subspace Search (e.g. HiCS[4])

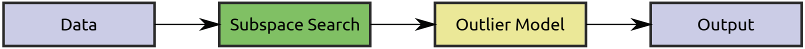


Related Work

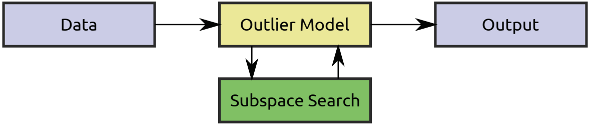
- Subspace Outlier Mining (e.g. OutRes[3])



- Subspace Search (e.g. HiCS[4])



- Flexible Subspace Search (RefOut[5])

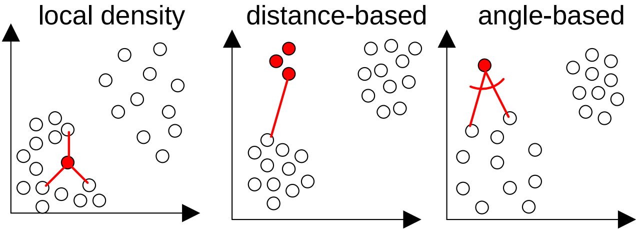


Flexible and Adaptive Subspace Search (RefOut)

- Outlier descriptions require an adaptive search w.r.t. outlier definition

⇒ Steer the search with some external objective function

$score_{LOF}(o, S) \rightarrow S_1$ $score_{NG}(o, S) \rightarrow S_2$ $score_{ABOF}(o, S) \rightarrow S_3$

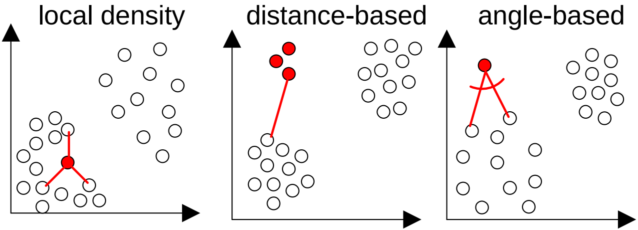


Flexible and Adaptive Subspace Search (RefOut)

- Outlier descriptions require an adaptive search w.r.t. outlier definition

⇒ Steer the search with some external objective function

$score_{LOF}(o, S) \rightarrow S_1 \qquad score_{NG}(o, S) \rightarrow S_2 \qquad score_{ABOF}(o, S) \rightarrow S_3$

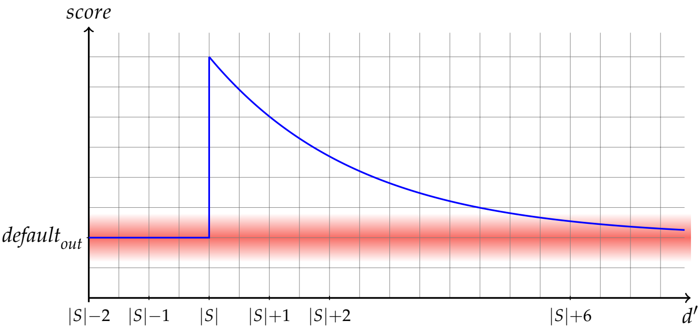


- RefOut is the first method that enables adaptive subspace search

⇒ It opens a new research direction: **subspace ensembles**[6]

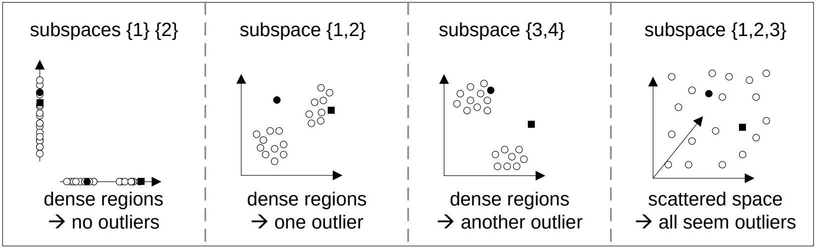
[6] Aggarwal: **Outlier ensembles: Position paper**, in ACM SIGKDD Explorations 2012.

Problem Setting in RefOut



- **For each object:**
  Search the **peak subspace** with best discriminative power
- Flexible search steered according to an external objective function

Problem Setting in RefOut

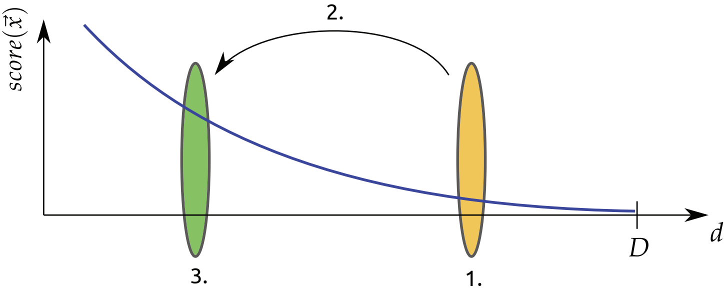


- **For each object:**
  Search the **peak subspace** with best discriminative power
- Flexible search steered according to an external objective function

RefOut Solution I

Algorithm

1. Apply outlier scoring to subspaces of an **initial subspace pool**
2. For the most promising outliers:
   Refine subspaces by identifying the peaking subspace
3. Apply outlier scoring to the **refined subspace pool**



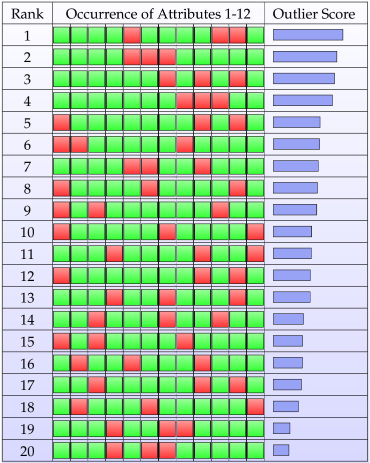
RefOut Solution II

- Given pool of subspaces
- Measure outlier score
  (or any other objective function)
- Combine best scoring subspaces

Score Discrepancy Problem:

Given a pool of subspaces and outlier scores, which subspace causes a partitioning $(\mathcal{O}_S^+, \mathcal{O}_S^-)$ that maximizes:

$$\arg\max_S (E[\mathcal{O}_S^+] - E[\mathcal{O}_S^-])$$



Rank	Occurrence of Attributes 1-12												Outlier Score
1	🟩	🟩	🟩	🟩	🟥	🟩	🟩	🟩	🟩	🟥	🟥	🟩	
2	🟩	🟩	🟩	🟩	🟥	🟥	🟥	🟩	🟩	🟩	🟩	🟩	
3	🟩	🟩	🟩	🟩	🟩	🟩	🟥	🟩	🟥	🟩	🟥	🟩	
4	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟥	🟥	🟥	🟩	🟩	
5	🟥	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟥	🟩	🟥	🟩	
6	🟥	🟥	🟩	🟩	🟩	🟩	🟩	🟥	🟩	🟩	🟩	🟩	
7	🟩	🟩	🟩	🟩	🟥	🟥	🟩	🟩	🟥	🟩	🟩	🟩	
8	🟥	🟩	🟩	🟩	🟩	🟥	🟩	🟩	🟩	🟩	🟥	🟩	
9	🟥	🟩	🟥	🟩	🟩	🟩	🟩	🟩	🟩	🟥	🟩	🟩	
10	🟥	🟩	🟩	🟩	🟩	🟩	🟥	🟩	🟩	🟩	🟩	🟥	
11	🟩	🟩	🟩	🟥	🟩	🟩	🟩	🟩	🟥	🟩	🟩	🟥	
12	🟥	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟥	🟩	🟥	🟩	
13	🟩	🟩	🟩	🟥	🟩	🟩	🟥	🟩	🟩	🟩	🟥	🟩	
14	🟩	🟩	🟥	🟩	🟩	🟩	🟥	🟩	🟩	🟥	🟩	🟩	
15	🟥	🟩	🟥	🟩	🟩	🟩	🟩	🟥	🟩	🟩	🟩	🟩	
16	🟩	🟥	🟩	🟩	🟥	🟩	🟩	🟩	🟥	🟩	🟩	🟩	
17	🟩	🟩	🟥	🟩	🟩	🟩	🟩	🟩	🟥	🟩	🟥	🟩	
18	🟩	🟥	🟩	🟩	🟩	🟥	🟩	🟩	🟩	🟩	🟩	🟥	
19	🟩	🟩	🟩	🟥	🟩	🟩	🟥	🟥	🟩	🟩	🟩	🟩	
20	🟩	🟩	🟩	🟥	🟩	🟥	🟥	🟩	🟩	🟩	🟩	🟩	

Evaluation, Application, and Extension

Evaluation

- Enhanced outlier detection quality (synthetic data)
- Provide meaningful outlier descriptions (real-world data)

Application of Subspace Search

- As multi-view feature selection
- As multi-view correlation analysis

Open for Academia and Industry

- Ensure repeatability of experiments by *OpenSubspace*[8]
- Provides outlier rules for outlier description[9]
- Extensible repository of algorithms (for academia and industry)

[8] Müller, Schiffer, Gerwert, Hannen, Jansen and Seidl: **SOREX: Subspace Outlier Ranking Exploration Toolkit**, in PKDD 2010.
[9] Müller, Keller, Blanc and Böhm: **OutRules: A Framework for Outlier Descriptions in Multiple Context Spaces**, in PKDD 2012.

Conclusion and Outlook

Subspace search is an emerging research field ...

Theoretical Models

- Statistical selection of relevant subspaces

⇒ How to exclude even mores undesired subspaces?

Algorithms

- Development of pruning heuristics

⇒ How to ensure scalability for large and complex data?

Descriptions

- Subspaces provide first descriptions

⇒ How to enable verification of patterns?