Ranking Outlier Nodes in Subspaces of Attributed Graphs

Emmanuel Müller, Patricia Iglesias, Yvonne Mülle, Klemens Böhm
Motivation

- Networks
  - Communication networks
  - Social networks
  - Auction networks
  - Co-purchased networks

- Application
  - Fraud detection
  - Spam detection
  - Network intrusion analysis
Example: Outlier Mining on Attributed Graphs

Input:

- **Node Attributes**
- **Graph Structure**

Output: Is a ranking of all nodes ordered by deviation w.r.t. subgraph and relevant attribute subspaces
Related Work: Outlier Mining

Vector Data


LOF (Breunig, et al.)

SOF (Aggarwal et al.)

Graph Data

SCAN (Xu et al.)

SUBDUE (Noble et al.)

Vector and Graph Data

CODA (Gao et al.)

GOutRank

Müller, Iglesias, Mülle, Böhm – Ranking Outlier Nodes in Subspaces of Attributed Graphs
(1) Selection of relevant subspaces and subgraphs

(2) Scoring of objects in multiple subspace clusters

(3) Availability of benchmark datasets
We propose a **decoupled process:**

1. **Selection:**
   - subgraphs
   - relevant subspaces

2. **Scoring:**
   - multiple subspace clusters
(1) Selection of Subspaces and Subgraphs

- Subspace clustering on attributed graphs
  - **Input:** graph \((V,E)\) and attributes \(A\)
  - **Output:** \(Res = \{(C_1, S_1) \ldots (C_n, S_n)\}\) with \(C_i \subseteq V\) and \(S_i \subseteq A\)

- Algorithmic solutions:
  - GAMer\(^{[1]}\)
  - Cocain\(^{[2]}\)
  - CoPam\(^{[3]}\)
  - ...

- Provide models for groups of similar nodes

---

Our GOutRank Framework

We propose a decoupled process:

- Database
- Subspace Clustering
- Subspace Clusters
- Outlier Scoring
- Ranking

How to derive an outlier score based on subspace cluster results?

Scoring:
- multiple subspace clusters
(2) Scoring with Multiple Subspace Clusters

- Properties of subspace clusters:
  - Overlap (i.e. objects belong to several clusters in different subspaces)
  - Different cluster sizes and dimensionality

Res:

\[
(C_1, S_1) = (\{o_3, o_4, o_5, o_7, o_8, o_9, o_{10}\}, \{d_1, d_2\})
\]
\[
(C_2, S_2) = (\{o_1, o_6, o_7, o_9, o_{10}, o_{11}, o_{12}, o_{13}, o_{14}\}, \{d_3\})
\]
\[
(C_3, S_4) = (\{o_2, o_5, o_9, o_{13}, o_{14}\}, \{d_1, d_2, d_4, d_5, d_6\})
\]

- Scoring function considering cluster properties\textsuperscript{[4]}

\[
score(o) = f(\text{Res})
\]

\textsuperscript{[4]} Müller et al.: “Outlier Ranking via Subspace Analysis in Multiple Views of the Data.” In IEEE ICDM 2012
### Combined Scored Function

- **Properties from the graph structure:**
  - Centrality of a node
  - Edge density of the subgraph (ongoing work)
  - Analysis of neighboring subspace clusters (ongoing work)

**Res:**

- \((C1, S1) = \{(o_3, o_4, o_5, o_7, o_8, o_9, o_{10}), \{d_1, d_2\}\}\)
- \((C2, S2) = \{(o_1, o_6, o_7, o_9, o_{10}, o_{11}, o_{12}, o_{13}, o_{14}), \{d_3\}\}\)
- \((C3, S4) = \{(o_2, o_5, o_9, o_{13}, o_{14}), \{d_1, d_2, d_4, d_5, d_6\}\}\)

**Graph:**

- Combine both sources of information:
  \[\text{score}(o) = f(\text{Res}, \text{Graph})\]
Experimental Setup

- Competitors
  - Only on vector data: full dimensional vs. subspace selection
  - Only on graph data: node outliers as by-product of graph clustering
  - On vector and graph data: community outlier detection

- Instantiation of different cluster models and scoring functions

- All experiments on:
  - subgraph of the Amazon co-purchase network
Outlier Identification

Setting of our user experiment
- **Users (high school students)**
  - No prior knowledge on outlier mining
  - Expertise by domain knowledge
- Attributed graph:
  - Disney DVDs (as Amazon products)
  - Presentation of co-purchased products (i.e. pre-computed graph clusters)

Tasks:
1. **Select outliers** in each set of co-purchased products
2. **Write an explanation** for the deviation of outliers
Our Benchmark Database

- Disney subgraph with **124 products, 334 edges**.
- Each product is labeled as outlier iff selected by >50% of the students

**Examples:**

- **Price:** 100$
  - **Suggested price:** 14,99$
  - (2003)

- High 1 Rating and low 5 Rating Ratio w.r.t. Pixar Films
# Evaluation w.r.t. Competitors

Comparison w.r.t. several outlier mining paradigms

<table>
<thead>
<tr>
<th>Database</th>
<th>Paradigm</th>
<th>Algorithm</th>
<th>AUC [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vector data</td>
<td>full data space</td>
<td>LOF[5]</td>
<td>56,85</td>
</tr>
<tr>
<td>Graph structure</td>
<td>graph clustering</td>
<td>SCAN[7]</td>
<td>52,68</td>
</tr>
<tr>
<td>Attributed Graph</td>
<td>full data space</td>
<td>CODA[8]</td>
<td>50,56</td>
</tr>
<tr>
<td></td>
<td>selected subspaces</td>
<td>GOutRank</td>
<td>86,86</td>
</tr>
</tbody>
</table>

Internal Evaluation

- Comparison of **Res** from different **subspace clustering models**
- Comparison of different **scoring functions**

<table>
<thead>
<tr>
<th>Res</th>
<th>Graph</th>
<th>AUC [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>GAMer(^1)</td>
<td>--</td>
<td>75.28</td>
</tr>
<tr>
<td></td>
<td>degree((o))</td>
<td>82.91</td>
</tr>
<tr>
<td></td>
<td>eigenvalue((o))</td>
<td>86.86</td>
</tr>
<tr>
<td>Extension of Cocain(^2)</td>
<td>--</td>
<td>75.85</td>
</tr>
<tr>
<td></td>
<td>degree((o))</td>
<td>76.97</td>
</tr>
<tr>
<td></td>
<td>eigenvalue((o))</td>
<td>77.96</td>
</tr>
<tr>
<td>CoPaM(^3)</td>
<td>--</td>
<td>58.61</td>
</tr>
<tr>
<td></td>
<td>degree((o))</td>
<td>69.49</td>
</tr>
<tr>
<td></td>
<td>eigenvalue((o))</td>
<td>72.45</td>
</tr>
</tbody>
</table>
## Conclusion & Outlook

- **Selection of subgraphs and subspaces**
  - Decoupled processing scheme exploiting subspace clusters

  - **Scalability to large attributed graphs**
  - Integration of outlier ranking into graph clustering algorithms

- **Scoring of objects in multiple subspace clusters**
  - Ranking combining graph structure and subspace cluster analysis

  - **Improvement of the scoring functions**
  - Extraction of more graph subspace cluster properties

- **Availability of benchmark datasets**
  - First benchmark on a subgraph from the Amazon co-purchased network

  - **Complete benchmark graph** (>300,000 nodes) with large user experiment (> 200 users)
Thank you for your attention

Our benchmark database is available online:

http://www.ipd.kit.edu/~muellere/GOutRank/