Reverse Nearest Neighbors in Unsupervised Distance-Based Outlier Detection*

* Article accepted in IEEE TKDE

Miloš Radovanović¹ Alexandros Nanopoulos²
Mirjana Ivanović¹

¹Department of Mathematics and Informatics
Faculty of Science, University of Novi Sad, Serbia

²Ingolstadt School of Management
University of Eichstaett-Ingolstadt, Germany
The Hubness Phenomenon

[Radovanović et al. ICML’09, Radovanović et al. JMLR’10]

- \( N_k(x) \), the number of \textbf{k-occurrences} of point \( x \in \mathbb{R}^d \), is the number of times \( x \) occurs among \( k \) nearest neighbors of all other points in a data set. In other words:
  - \( N_k(x) \) is the reverse \( k \)-nearest neighbor count of \( x \)
  - \( N_k(x) \) is the in-degree of node \( x \) in the \( k \)NN digraph

- Observed that the distribution of \( N_k \) can become skewed, and have high variance, resulting in \textbf{hubs} – \textbf{points with high} \( N_k \) \textbf{values}, and \textbf{anti-hubs} – \textbf{points with low} \( N_k \)
  - Music retrieval [Aucouturier & Pachet PR’07]
  - Speaker verification (“Doddington zoo”) [Doddington et al. ICSLP’98]
  - Fingerprint identification [Hicklin et al. NIST’05]

- Cause remained unknown, attributed to the specifics of data or algorithms
iid uniform, $d = 3$, $corr = -0.020$

iid uniform, $d = 20$, $corr = -0.800$

iid uniform, $d = 100$, $corr = -0.865$
Causes of Hubness

- **Related phenomenon: concentration of distance / similarity**
  - High-dimensional data points approximately lie on a sphere centered at any fixed point [Beyer et al. ICDT’99, Aggarwal & Yu SIGMOD’01]
  - The distribution of distances to a fixed point always has non-negligible variance [François et al. TKDE’07]
  - As the fixed point we observe the data set center

- **Centrality**: points closer to the data set center tend to be closer to all other points (regardless of dimensionality)
  
  *Centrality is amplified by high dimensionality*
Important to Emphasize

- Generally speaking, concentration does not CAUSE hubness

- “Causation” might be possible to derive under certain assumptions. My preferred view: they are both manifestations of underlying mechanisms triggered by high dimensionality

- Example settings with(out) concentration and with(out) hubness:
  - C+, H+: iid uniform data, Euclidean dist.
  - C–, H+: iid uniform data, squared Euclidean dist.
  - C+, H–: iid normal data (centered at 0), cosine sim.
  - C–, H–: spatial Poisson process data, Euclidean dist.

- Two “ingredients” needed for hubness:
  1) High dimensionality
  2) Centrality (existence of centers / borders)
Hubness in Real Data

- Important factors for real data
  1) Dependent attributes
  2) Grouping (clustering)

- 50 data sets
  - From well known repositories (UCI, Kent Ridge)
  - Euclidean and cosine, as appropriate

- Conclusions [Radovanović et al. JMLR’10]:
  1) Hubness depends on intrinsic dimensionality
  2) Hubs are in proximity of cluster centers
Anti-Hubs in Outlier Detection

[Radovanović et al. JMLR’10]

- In high dimensions, points with low $N_k$ – the anti-hubs can be considered **distance-based outliers**
  - They are far away from other points in the data set / their cluster
  - High dimensionality contributes to their existence

![Graphs: Ionosphere and Sonar datasets](NII, Tokyo)

March 23, 2015
Anti-Hubs in Outlier Detection

[Aggarwal and Yu SIGMOD’01]
● In high-dimensional space unsupervised methods detect every point as an almost equally good outlier, since distances become indiscernible as dimensionality increases

[Zimek et al. SADM’12]
● The above view was challenged by showing that the exact opposite may take place
● As dimensionality increases, outliers generated by a different mechanism from the data tend to be detected as more prominent by unsupervised methods
  ○ Assuming all dimensions carry useful information
Anti-Hubs in Outlier Detection

- We show that the opposite can take place even when no true outliers exist, in the sense of originating from a different distribution.
- This suggests that high dimensionality affects outlier scores and (anti-)hubness in similar ways.
Hubness and Large Neighborhoods

$d = 3$

$d = 20$

$d = 100$

March 23, 2015
Hubness and Large Neighborhoods

iid uniform, $d = 3$, $corr = -0.999$

iid uniform, $d = 20$, $corr = -0.999$

iid uniform, $d = 100$, $corr = -0.999$
Hubness and Large Neighborhoods

- $p$ = percentage of points with lowest $N_k$ scores
- High dimensionality ($d$): $N_k$ strong indicator of centrality overall ($p = 100\%$), but weaker for anti-hubs ($p = 5\%$)
- Low $d$: the opposite, especially w.r.t low $k$ values
- Raising $k$ strengthens correlation, but not when cluster boundary is crossed
The AntiHub Method

[Hautamäki et al. ICPR’04]

- Proposed method ODIN (Outlier Detection using Indegree Number), which selects as outliers points with $N_k$ below or equal to a user-specified threshold
- Experiments on 5 data sets showed it can work better than various $k$NN distance methods
- Not aware of the hubness phenomenon, little insight into reasons why ODIN should work, its strengths, weaknesses…
- In method AntiHub, we use $N_k(x)$ as the outlier score of $x$ (same as ODIN, without the threshold)
The AntiHub Method

Algorithm 1 AntiHub_{dist}(D, k) (based on ODIN [11])

Input:
- Distance measure \( dist \)
- Ordered data set \( D = (x_1, x_2, \ldots, x_n) \), where \( x_i \in \mathbb{R}^d \), for \( i \in \{1, 2, \ldots, n\} \)
- No. of neighbors \( k \in \{1, 2, \ldots\} \)

Output:
- Vector \( s = (s_1, s_2, \ldots, s_n) \in \mathbb{R}^n \), where \( s_i \) is the outlier score of \( x_i \), for \( i \in \{1, 2, \ldots, n\} \)

Temporary variables:
- \( t \in \mathbb{R} \)

Steps:
1) For each \( i \in (1, 2, \ldots, n) \)
2) \( t := N_k(x_i) \) computed w.r.t. \( dist \) and data set \( D \setminus x_i \)
3) \( s_i := f(t) \), where \( f : \mathbb{R} \to \mathbb{R} \) is a monotone function
The AntiHub Method

- We experimentally identified strengths and weaknesses of AntiHub with respect to different properties (factors):
  1. Hubness
  2. Locality vs. globality
  3. Discreteness of scores
  4. Varying density
  5. Computational complexity
The AntiHub Method

Property 1: Hubness

- High (intrinsic) dimensionality, $k << n$:
  - Good overall correlation between $N_k$ and distance to a center, but
  - Many $N_k$ values of 0 – problem with discrimination

- Low dimensionality, $k << n$
  - Low correlation between $N_k$ and distance to a center, but
  - For a small number of points with low $N_k$, this correlation is better, so AntiHub/ODIN can be meaningful
The AntiHub Method

iid uniform, $N_k$: Dist. from data set mean

2c uniform, $N_k$: Dist. from cluster mean
The AntiHub Method

Property 2: Locality vs. globality
- For AntiHub and other methods based on $k$NN:
  - $k << n$: notion of outlierness is local
  - $k \sim n$: notion of outlierness is global
- AntiHub in “local mode” may have problems with discrimination
- Raising $k$ can address this, but the notion of outlierness goes global
  - This can be problematic if we are interested in local outliers, but $k$ crosses cluster boundaries

Property 3: Discreteness of scores
- Regardless of all of the above, $N_k$ scores are integers, hence inherently discrete, which can also cause discrimination problems
The AntiHub Method

Property 4: Varying density
- AntiHub is not sensitive to the scale of distances in the data
- Can effectively detect (local) outliers in clusters of different densities without explicitly modeling density

Property 5: Computational complexity
- Using high $k$ values can be useful
- However, approximate $k$NN search/indexing methods typically assume $k = O(1)$
The AntiHub² Method

- Notable weakness of AntiHub, discrimination of scores, contributed to by two factors:
  - Hubness
  - Discreteness of scores

- Therefore, we proposed method AntiHub², which combines the $N_k$ score of a point with $N_k$ scores of it’s $k$ nearest neighbors, in order to maximize discrimination

- AntiHub² improves discrimination of scores compared to the AntiHub method
The AntiHub^2 Method

Algorithm 2 AntiHub^2_{dist} (D, k, p, step)

Input:
- Distance measure dist
- Ordered data set \( D = (x_1, x_2, \ldots, x_n) \), where \( x_i \in \mathbb{R}^d \), for \( i \in \{1, 2, \ldots, n\} \)
- No. of neighbors \( k \in \{1, 2, \ldots\} \)
- Ratio of outliers to maximize discrimination \( p \in (0, 1] \)
- Search parameter \( step \in (0, 1] \)

Output:
- Vector \( s = (s_1, s_2, \ldots, s_n) \in \mathbb{R}^n \), where \( s_i \) is the outlier score of \( x_i \), for \( i \in \{1, 2, \ldots, n\} \)

Temporary variables:
- AntiHub scores \( a \in \mathbb{R}^n \)
- Sums of nearest neighbors’ AntiHub scores \( a_{nn} \in \mathbb{R}^n \)
- Proportion \( \alpha \in [0, 1] \)
- (Current) discrimination score \( c_{disc}, disc \in \mathbb{R} \)
- (Current) raw outlier scores \( c_t, t \in \mathbb{R}^n \)
The AntiHub² Method

Local functions:
- `discScore(y, p)`: for $y \in \mathbb{R}^n$ and $p \in (0, 1]$ outputs the number of unique items among $\lfloor np \rfloor$ smallest members of $y$, divided by $\lfloor np \rfloor$

Steps:
1) $a := \text{AntiHub}_{dist}(D, k)$
2) For each $i \in (1, 2, \ldots, n)$
3) $ann_i := \sum_{j \in \text{NN}_{dist}(k, i)} a_j$, where $\text{NN}_{dist}(k, i)$ is the set of indices of $k$ nearest neighbors of $x_i$
4) $disc := 0$
5) For each $\alpha \in (0, \text{step}, 2 \cdot \text{step}, \ldots, 1)$
   5) For each $i \in (1, 2, \ldots, n)$
6) $ct_i := (1 - \alpha) \cdot a_i + \alpha \cdot ann_i$
7) $cdisc := \text{discScore}(ct, p)$
8) If $cdisc > disc$
   9) $t := ct$, $disc := cdisc$
10) For each $i \in (1, 2, \ldots, n)$
11) $s_i := f(t_i)$, where $f : \mathbb{R} \to \mathbb{R}$ is a monotone function
Discrimination Improvement

discScore values for real data ($\rho = 10\%$, $\text{step} = 0.01$)
Performance Evaluation

Methods for comparison:

- $k$NN: distance to the $k$th nearest neighbor
  [Ramaswamy et al. SIGMOD Rec’00]

- ABOD: Angle Based Outlier Detection
  [Kriegel et al. KDD’08]

- LOF: Local Outlier Factor
  [Breunig et al. SIGMOD Rec’00]

- INFLO: INFLuenced Outlierness
  [Jin et al. PAKDD’06]
Performance Evaluation

- **Synthetic data:** two well-separated Gaussian clusters of the same size, std of one 10 times larger than other, outliers 5% of points from each cluster projected 20% farther from respective cluster center.
Performance Evaluation

- **Real data:** mostly natural labeled outliers from various domains

<table>
<thead>
<tr>
<th>Name</th>
<th>$n$</th>
<th>$d$</th>
<th>$S_{N_{10}}$</th>
<th>Outlier%</th>
</tr>
</thead>
<tbody>
<tr>
<td>aloi</td>
<td>50,000</td>
<td>64</td>
<td>0.260</td>
<td>3.016</td>
</tr>
<tr>
<td>churn</td>
<td>5,000</td>
<td>17</td>
<td>0.849</td>
<td>14.140</td>
</tr>
<tr>
<td>ctg3</td>
<td>2,126</td>
<td>35</td>
<td>0.652</td>
<td>8.279</td>
</tr>
<tr>
<td>ctg10</td>
<td>2,126</td>
<td>35</td>
<td>0.652</td>
<td>2.493</td>
</tr>
<tr>
<td>kdd99-r2l</td>
<td>68,338</td>
<td>38</td>
<td>0.018</td>
<td>1.456</td>
</tr>
<tr>
<td>kdd99-u2r</td>
<td>67,395</td>
<td>38</td>
<td>0.031</td>
<td>0.077</td>
</tr>
<tr>
<td>mammography</td>
<td>11,183</td>
<td>6</td>
<td>0.103</td>
<td>2.325</td>
</tr>
<tr>
<td>nba-allstar-1951-1972</td>
<td>4,018</td>
<td>15</td>
<td>0.483</td>
<td>15.903</td>
</tr>
<tr>
<td>nba-allstar-1973-2009</td>
<td>16,916</td>
<td>17</td>
<td>0.730</td>
<td>5.669</td>
</tr>
<tr>
<td>thyroid-sick</td>
<td>3,772</td>
<td>52</td>
<td>0.371</td>
<td>6.124</td>
</tr>
<tr>
<td>us-crime</td>
<td>1,994</td>
<td>100</td>
<td>1.327</td>
<td>7.523</td>
</tr>
<tr>
<td>wilt</td>
<td>4,839</td>
<td>5</td>
<td>-0.075</td>
<td>5.394</td>
</tr>
</tbody>
</table>
Performance Evaluation

- Two types of data sets: mostly local and mostly global outliers

- With respect to different $k$ values, AUC of AntiHub and AntiHub$^2$ behaves similarly to density-based methods (LOF, INFLO)

- Very high $k$ values can be useful for all methods, especially LOF, INFLO, AntiHub and AntiHub$^2$, suggesting there may be a relationship between “global” density-based and distance-based outliers

- AntiHub$^2$ can improve AUC of AntiHub, but not always, thus discrimination is not the only factor that should be addressed
Conclusions

- We provided a unifying view of the role of reverse nearest neighbor counts in unsupervised outlier detection:
  - Effects of high dimensionality on unsupervised outlier-detection methods and hubness
  - Extension of previous examinations of (anti-)hubness to large values of $k$
  - The article also explores the relationship between hubness and data sparsity
- We formulated the AntiHub method, discussed its properties, and improved it in AntiHub$^2$ by focusing on discrimination of scores
- Our main hope: clearing the picture of the interplay between types of outliers and properties of data, filling a gap in understanding which may have so far hindered the widespread use of reverse neighbor methods in unsupervised outlier detection
Future Possibilities

- High values of $k$ can be useful, but:
  - Cluster boundaries can be crossed, producing meaningless results of local outlier detection. How to determine optimal neighborhood size(s)?
  - Computational complexity is raised; approximate NN search/indexing methods do not work any more. Is it possible to solve this for large $k$?

- AntiHub and AntiHub$^2$ are no “rock star” methods
  - Can $N_k$ scores be applied to outlier detection in a better way? Through outlier ensembles?

- Extend to (semi-)supervised outlier detection methods
Future Possibilities

- Explore relationships between intrinsic dimensionality, distance concentration, (anti-)hubness, and their impact on subspace methods for outlier detection

- Investigate secondary measures of distance/similarity, such as shared-neighbor distances
References

M. Radovanović et al. Reverse nearest neighbors in unsupervised distance-based outlier detection. IEEE Transactions on Knowledge and Data Engineering, 2015 (forthcoming).
A. Hicklin et al. The myth of goats: How many people have fingerprints that are hard to match? Internal Report 7271, National Institute of Standards and Technology (NIST), USA, 2005.


