What is fraud detection?

- Identifying wrongful acts performed by the wrong people

Possible solution to detect fraud

- Decision tree and rules
  - Build a profile of the characteristics of fraudulent behavior
  - Pull out cases that meet those characteristics
- Clustering and identifying outliers
  - Group behavior using clustering algorithms
  - Find outliers and then investigate
- False positives

What is an outlier?

- An outlier is an observation that deviates so much from other observations as to arouse suspicions that it was generated by a different mechanism.
- Simple example: Stolen credit card
  - Card owner is concerned with implications of purchases and thief is not
  - Both have different likes/dislikes, shopping habits, and locations

Algorithm NL

- Complexity of $O(kN^2)$
- Algorithm tries to define a DB (Distance Based) outlier as: an object $O$ in a data-set $T$ is a DB($p$,$D$) outlier if at least fraction $p$ of the objects in $T$ lies greater than distance $D$ from $O$.

Definition 1

DB($p$,$D$) unifies1 or generalizes another definition Def for outliers, if there exist specific values $p_0, D_0$ such that: object $O$ is an outlier according to Def iff $O$ is a DB($p_0; D_0$) outlier.

Definition 2

Let $T$ be a normally distributed random variable with mean $\mu$ and standard deviation $\sigma$. Def$_{\text{Normal}}$ as follows: $t \in T$ is an outlier iff $(t - \mu) / \sigma \geq 3$ or $(t - \mu) / \sigma \leq -3$.

For a normal distribution, outliers can be considered to be points that lie three or more standard deviations from the mean.
Pseudo-code

1. Fill the first array with a block of tuples from T.

2. For each tuple $t_i$ in the first array, do:
   a. $count_i \leftarrow 0$
   b. For each tuple $t_j$ in the first array, if $\text{dist}(t_i, t_j) \leq D$ do:
      Increment $count_i$ by 1. If $count_i > M$, mark $t_i$ as a non-outlier and proceed to next $t_i$.

3. While blocks remain to be compared to the first array, do:
   a. Fill the second array with another block.
   b. For each unmarked tuple $t_i$ in the first array do:
      For each tuple $t_j$ in the second array, if $\text{dist}(t_i, t_j) \leq D$ do:
      Increment $count_i$ by 1. If $count_i > M$ do:
      mark $t_i$ as a non-outlier and proceed to next $t_i$.

4. For each unmarked tuple $t_i$ in the first array, report $t_i$ as an outlier.

5. If the second array has served as the first array anytime before, stop; otherwise, swap the names of the first and second arrays and go-to step 2.

Improvements

Fill the arrays with the blocks in the following order and comparing:

<table>
<thead>
<tr>
<th>Block Comparisons</th>
</tr>
</thead>
<tbody>
<tr>
<td>A,A</td>
</tr>
<tr>
<td>A,B</td>
</tr>
<tr>
<td>A,C</td>
</tr>
<tr>
<td>A,D</td>
</tr>
</tbody>
</table>

| 4 block reads      | 2 block reads      | 2 block reads      | 2 block reads      |

Total block reads: 10

Number of passes over the entire data set: $10/4 = 2.5$

References