Efficient Large-Scale Local Outlier Detection

Timothy de Vries *Supervisor: Sanjay Chawla*

School of Information Technologies

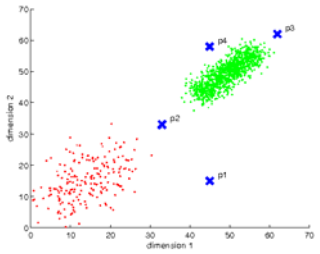
FACULTY OF ENGINEERING & INFORMATION TECHNOLOGIES

What are local anomalies?

Local anomalies or outliers are points that are anomalous with respect to their immediate neighbourhood instead of the data set as a whole. Examples:

- Credit card fraud
- Network intrusion
- Other abnormal behaviour in a local region (e.g. stock market time period)
- Cancer cells
- Image data with occlusion and shadows
- Free text anomalies

Examples: blue X in graph below are *locally* distinct but not *globally* (some red points are bigger global outliers)



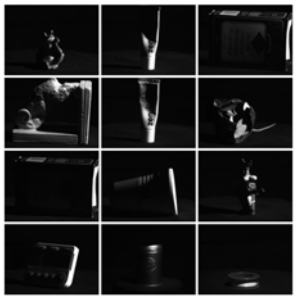
Standard (global) outlier example

From real world data set of object images

- 24,000 pictures of 1000 objects, 24 pictures per object, varying the direction of lighting
- Images are192x144 pixels = 27648 ambient dimensionality, but low intrinsic dimensionality
- Top Global outliers correspond to distinct and bright shapes that are most different from every other object in the data set

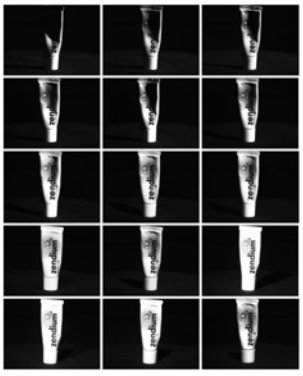
Local outlier example

- The 24 pictures of each object form a cluster in the data set
- When a picture is changed substantially via lighting changes or the appearance of shadows, the picture may move a short distance away from the cluster, forming a local outlier
- The top local outliers of the whole data set are among the most occluded by shadows and low illumination – difficult cases for image recognition



Practical example

- The set of pictures for a single object (toothpaste), when ordered by Local Outlier Factor, allows us to measure the level of distortion via illumination changes and shadow occlusion
- This information, along with the nearest-neighbours found using Local Outlier Factor, allow us to easily classify new images as being a possibly occluded image of an existing item, or a new item altogether

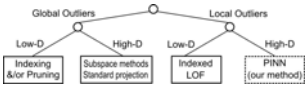


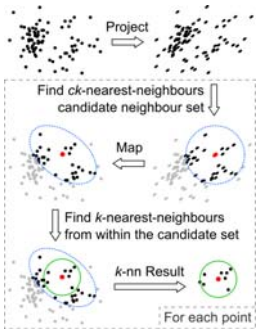
The Algorithm (PINN)

Allows us to find local outliers in very large and high-dimensional real data

We use dimensionality reduction to project data to a lower dimension

- Allows efficient indexing
- Near-linear algorithm, improvement from O(n2)
- Distances between points are preserved through projection
- First known method for high-dimensional local outliers:





$$\overline{LOF}(p) = \frac{\frac{1}{k}\sum_{q\in\overline{N}_k(p)} rd(q)}{rd(p)}.$$

$$\frac{1-\epsilon}{1+\epsilon}\cdot LOF(p) \le \overline{LOF}(p) \le \frac{1+\epsilon}{1-\epsilon}\cdot LOF(p).$$

Algorithm 1 RP + PINN + LOF

Input: The n by m matrix of data in the original space, denoted X.

Output: The Local Outlier Factor (LOF) score and ranking for each point in X.

RP:

Project X to a n by t matrix Y, $t < m$, using the random projection scheme described in Section II-B.

PINN:

Define h as the parameter for defining the size of the set of candidate nearest-neighbours used, where $h \geq k$.

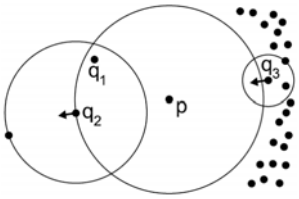
For each point $p \in X$:

1) Find h-nearest-neighbours of p' in the projected space Y, forming the candidate nearest-neighbour set $N_h(p')$.
2) Map the points in the candidate set $N_h(p')$ back to the original space (X), forming the set $\mathrm{RP}^{-1}(N_h(p'))$.
3) Find the k items of $\mathrm{RP}^{-1}(N_h(p'))$ closest to p. Call this set $\overline{N}_k(p)$

LOF:

For each point $p \in X$, estimate $LOF(p)$ by computing

$$\overline{LOF}(p) = \frac{\frac{1}{k}\sum_{q\in\overline{N}_k(p)} rd(q)}{rd(p)}.$$



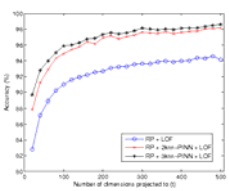
Solution to the neighbour-replacement problem that occurs in dimensionality reduction or projection for local outlier detection using Local Outlier Factor

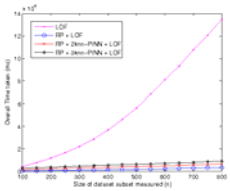
Results

Accuracy and performance on real world data sets from a wide range of different type (Image, Text, Medical, Mixed)
Red and black lines refer to the new method (with parameter changes). Pink is the standard approach without optimisation
Blue is the baseline standard approach Random Projection. Green is baseline Principal Components Analysis.

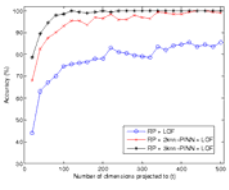
❑ Accuracy ❑ Performance

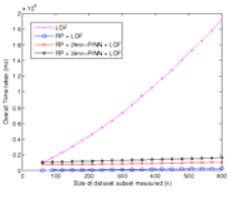
❑ Yale face image data set





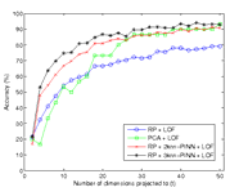
❑ Internet advert detection

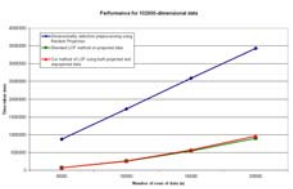




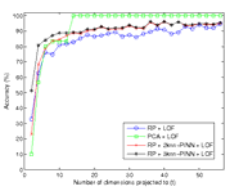
❑ Accuracy ❑ Performance

❑ Medical data (colon cancer)





❑ Spam detection



❑ Performance on the large-scale Reuters New York Times text article data set (for text mining)
- ❑ 300,000 rows
- ❑ 102,600 dimensions

THIS RESEARCH IS SPONSORED BY

Capital Markets CRC Limited