Poster Authors:
Peter Aláč
Simon Sudora

Masaryk University, Brno

Local Outlier Detection with Interpretation

Xuan Hong Dang[1], Barbora Micenkova[1], Ira Assent[1], and Raymond T. Ng[2]
[1] ***Aarhus University, Denmark***
{dang,barbora,ira}@cs.au.dk
[2] ***University of British Columbia, Canada***
rng@cs.ubc.ca

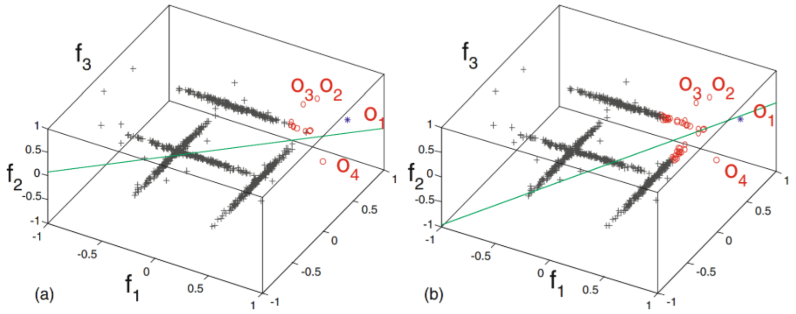
INTRODUCTION

Outlier detection aims at searching for a small set of objects that are inconsistent or considerably deviating from other objects in a dataset. Existing research focuses on outlier identification while omit-ting the equally important problem of outlier interpretation. This poster presents a novel method of local outlier detection named LODI to address both problems at the same time.

NEIGHBORING SET SELECTION

Most common approach is using the set of k nearest neighboring objects (kNNs). The problems of this approach:

- Identifying a proper value of k is non-trivial task.
- Neighbors might also contain nearby outliers or inliers from several distributions.

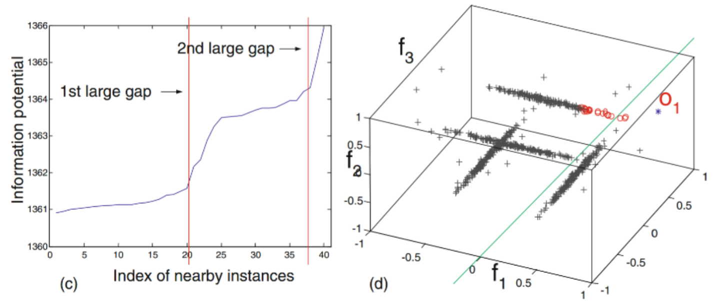


We use an adaptive technique based on the concept of **entropy in information theory**, which doesn't fix the number of neighboring inliers k. Entropy measures the uncertainty. Shannon's definition of entropy: $H(X) = -\int p(\mathbf{x}) \log p(\mathbf{x}) d\mathbf{x}$

Computing entropy in Shannon's definition is not an easy task since it requires p(x) to be known. We need more general formula. Combining Renyi entropy and Parzen window technique for estimation p(x) we get **local quadratic Renyi entropy** as follows: $QE(R(\mathbf{o})) = -\ln \frac{1}{s^2} \sum_i^s \sum_j^s G(\mathbf{x}_i - \mathbf{x}_j, 2\sigma^2)$

Each term in the summation increases as the distance between xi and xj decreases. **Intuition**: The higher the information potential of the set is, the more similar the elements within the set are. **Objective**: Maximizing the information potential within the neighboring set is equivalent to minimizing the entropy. Naive way to find an optimal set of neighbors may require computing all possible combinations, which is too expensive. Therefore, we use a heuristic approach.

Intuition: Removing an object from the neighboring set will lead to a decrement in the total information potential. Those instances resulting in the most decrement are important ones and vice versa. Proposed **heuristic method** ranks the total information potential left in the increasing order and removes objects behind the first significant gap (larger than the average gap).



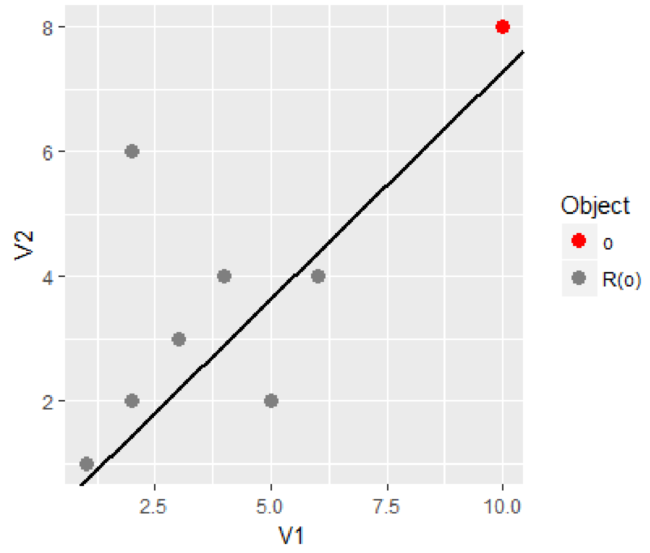
DIMENSIONALITY REDUCTION

Given the calculated neighboring set, the next step is to calculate the anomaly degree for each object in the dataset. In order to do so, we at first perform a local dimensionality reduction. In particular, we identify a subspace **w**, in which:

•candidate outlier **o** is strongly deviating from every object in its neighborhood R(**o**) e.g. it has high distance D(**o**, R(**o**))

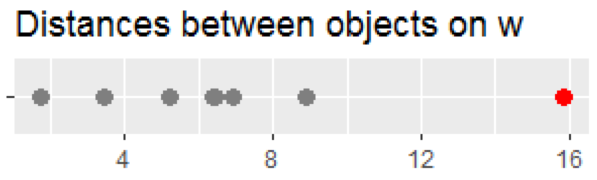
•while at the same time R(**o**) shows low variance Var(R(o)) in that induced subspace.

We use eigen-decomposition in order to find such a vector which maximize the ratio D(**o**, R(**o**)) / Var(R(**o**)). Here is an example of such a vector w in 2-dimensional space (the objects have only 2 features V1, V2):

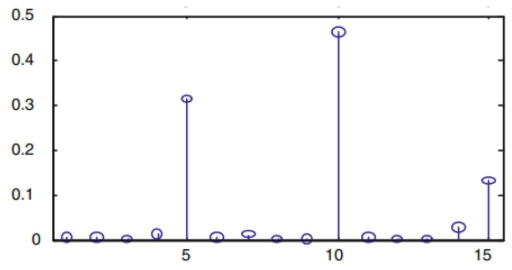


OUTLIER INTERPRETATION

Obtained eigenvector **w** represents (always) 1-dimensional subspace. Using **w** we can simply measure the distance between **o** and each R(**o**).

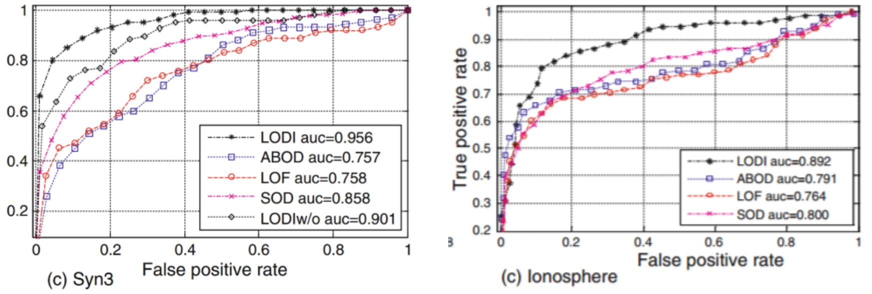


The local outlier is identified when its average distance is higher than the average distances of its neighbors. Therefore relative distance for the rightmost point, for instance, would be high. We denote this measure as a local anomaly degree. We can order objects based on their local anomaly degree to identify, which are most likely local outliers. Furthermore, we can use w to interpret, which feature is the most important in determining o as an outlier. Coefficients of the eigenvector w are truly the weights of the original features. Here is a different example of feature weights of an object, where features 10, 5 and 15 have the biggest impact on classifying an object as an outlier:



EXPERIMENTAL RESULTS

- Syn3 – synthetic dataset with 50k instances and 50 dimensions.
- Ionosphere – real world dataset with 351 instances and 32 dimensions



Note: LODIw/o - kNNs selection used.