A New Local Distance-based Outlier Detection Approach for Scattered Real-World Data

Ke Zhang[1], Marcus Hutter[1,2] and Huidong Jin[1,2,3]

[1]RSISE, Australian National University,

[2]NICTA, Canberra Lab, ACT, Australia,

[3]CSIRO Mathematical and Information Sciences, ACT, Australia.

April 24, 2009

Outlier detection

- Outlier: an observation (or measurement) that is unusually different (large or small) from others in a dataset.
- Causes:
    - record or measurement error;
    - contamination from different data population;
    - inherent variability, e.g. rare event.
- Application:
    - medical (e.g. adverse reactions analysis),
    - finance (e.g. financial fraud detection),
    - security (e.g. counter-terrorism),
    - information security (e.g. intrusions detection).

Challenges in real-world application

Challenges in real-world applications:

- parameter setting problem;
  - no outlier labels
  - can not optimise their parameters through trail-and-error
- scattered data structure, which does not explicitly represent normal data 'behaviors'.

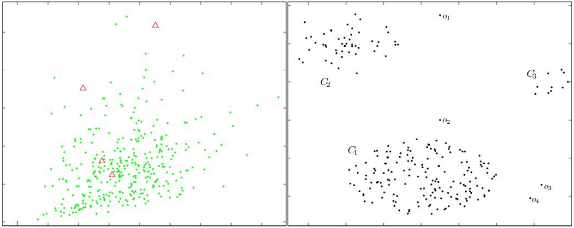
Top-n outlier

With the consideration of the parameter setting problem, researchers proposed top-n style outlier detection methods.

- only have one crucial parameter, less than other OD methods;
- short-list the n most suspicious objects with the highest 'outlier-ness' factor;
- provide a good interaction between technique provider and user;
- typical methods: top-n KNN top-n LOF.

Problems in scattered data

- Objects are scattered distributed in feature space.
- Locally, objects are randomly allocated.
- Globally, the scattered objects constitute lots of mini-clusters.
- Outlier: the object deviating from any other group.



(a) (b)

Figure: (a) The 2-D projection of a real-world dataset. (b) Simple 2-D illustration

Local distance-based outlier factor definition

The k-nearest neighbours distance of object x_p is defined as

$$\bar{d}_{x_p} := \frac{1}{k} \sum_{x_i \in \mathcal{N}_p} \mathrm{dist}(x_i, x_p).$$

The k-nearest neighbours inner distance of x_p :

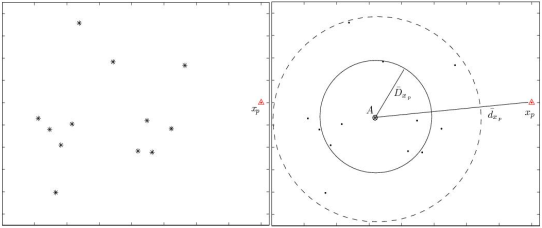
$$\bar{D}_{x_p} := \frac{1}{k(k-1)} \sum_{x_i, x_{i'} \in \mathcal{N}_p, i \neq i'} \mathrm{dist}(x_i, x_{i'}).$$

The local distance-based outlier factor of x_p is defined as:

$$LDOF_k(x_p) := \frac{\bar{d}_{x_p}}{\bar{D}_{x_p}}$$

Local distance-based outlier factor

- LDOF uses the relative position of an object to its neighbours to indicate the degree of the object deviating from its neighbourhood system.
- The k-nearest neighbours distance of x_p equals the average distance from x_p to all objects in $\mathcal{N}_p$.
- The k-nearest neighbours inner distance of x_p is defined as the average distance among objects in $\mathcal{N}_p$.



(a) (b)

LDOF properties

- Let data $\mathcal{D}$ be uniformly distributed in a neighbourhood of x_p containing k objects. For large k, we have $LDOF_{lb} \approx \frac{1}{2}$ with high probability.
- Let data $\mathcal{D}$ be uniformly distributed in a neighbourhood of x_i containing k objects. For $LDOF > \frac{1}{2}$, the probability of false detecting $x_p \in \boldsymbol{R}^d$ as an outlier is exponentially small in k. More precisely, the probability of false-detection is:

$$P_{\text{false-detection}} \; < \; e^{-\alpha(k-2)}, \quad \text{where} \quad \alpha := \frac{2}{25}(1-\frac{1}{2LDOF})^2(\frac{d}{d+2})^2$$

Top-n LDOF

Top-n local distance-based outlier detection approach:

1. **Input:** A given dataset $\mathcal{D}$, natural numbers n and k.
2. For each object p in $\mathcal{D}$, retrieve p's k-nearest neighbours;
3. Calculate the *LDOF* for each object p. The objects with $LDOF < LDOF_{lb}$ are directly discarded;
4. Sort the objects according to their *LDOF* values;
5. **Output:** the first n objects with the highest *LDOF* values.

Synthetic dataset

Experimental results:

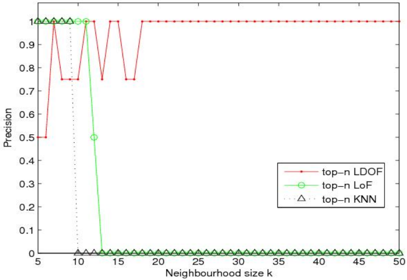
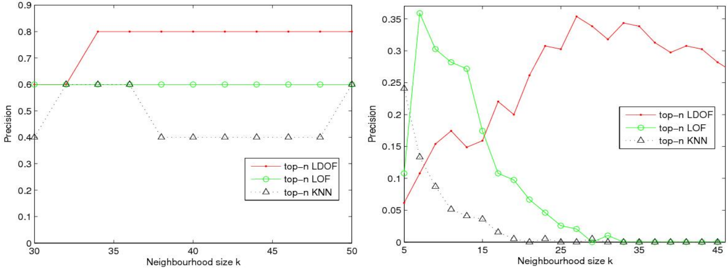


Figure: Detecting precisions of top-n LDOF, top-n KNN, top-n LOF on a synthetic dataset.

Real-world data



(a) Medical diagnosis dataset. (b) Space shuttle dataset.

Figure: Detecting precisions of top-n LDOF, top-n KNN, top-n LOF on real-world datasets.

Conclusion

- Proposed a local distance-based outlier factor for solving scattered data problem.
- Employed top-n technique to facilitate parameter setting.
- Suggested the method of selecting neighbourhood size k.
- Demonstrated the ability of LDOF to better discover outliers (high precision, stable over a large range of k).
- Future work: further enhance the outlier detection accuracy for scattered real-world datasets.

Thank You!