Discovering NULL & Outliers

404 Not Found

Guanhua Chen

Yicong Xu

Wei Wang

Outline

1. Introduction
2. Data preprocessing
   a. One hot encoding
   b. Normalization
   c. Null detection
3. Methods:
   a. Distance-based
      i. Nestedloop
      ii. Index-based kd-tree
      iii. Innovation -- LOF
   b. Cluster-based
      i. K-Means
      ii. DBSCAN
4. Results
5. Reference & Codes

Normalization

Example column
1
2
2
4

range:(1,4)

After nomalization
0
(2-1)/3=1/3
(2-1)/3=1/3
1

Null detection

Use spark to find the NULL value

Too many NULL in one row

Delete the row

Very few NULL values in one row

Treat the NULL value as a special value

Encoding

Fruit
Apple
Orange
Orange
Pear

Origin column	After encoding
Apple	1
Orange	2
Orange	2
Pear	3

Label encoding

Origin column	Apple	Orange	Pear
Apple	1	0	0
Orange	0	1	0
Orange	0	1	0
Pear	0	0	1

One hot encoding

Distance-based Method

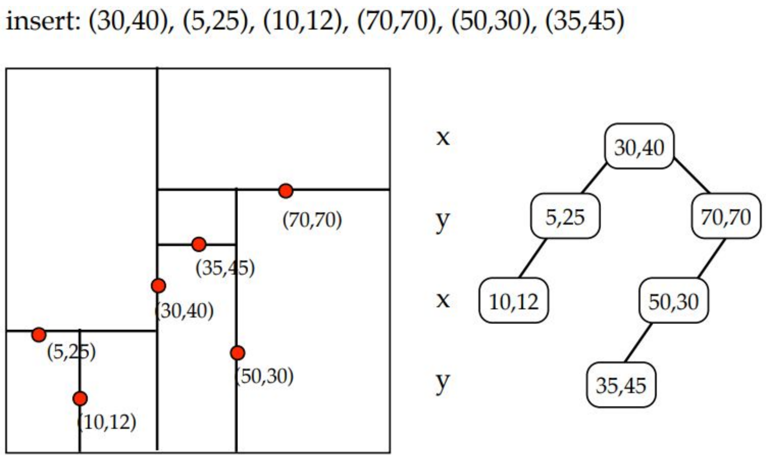
- General Idea
  - Judge a point based on the distance(s) to its neighbors
  - Several variants proposed
- Basic Assumption
  - Normal data objects have a dense neighborhood
  - Outliers are far apart from their neighbors, i.e., have a less dense neighborhood

Kth-Neighbor Algorithm

- General models
  - Take the KNN distance of a point as its outlier score [Ramaswamy et al 2000]
  - The larger KNN distance -> Sparser neighborhood -> The point far apart from its neighborhood -> Outlier
- Algorithm:
  - Nest-Loop(Naïve approach):
    - For each object: compute kNNs with a sequential scan
    - Rank the socres: Higher score -> high Outlier Possibility
    - Complexity: O(δN^2)

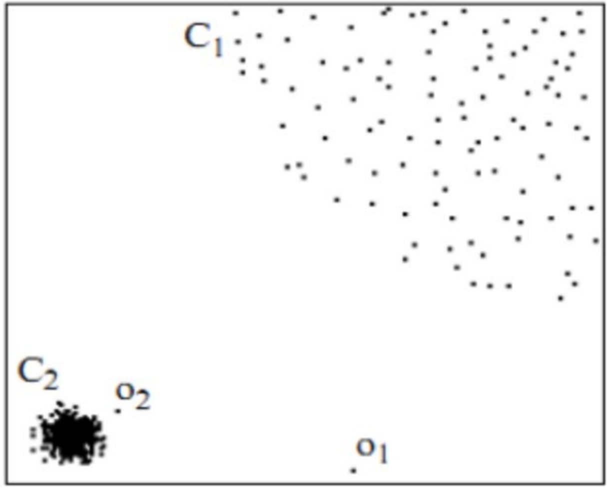
Kth-Neighbor Algorithm--Using KD-Tree

- Algorithm:
  - Use KD-Tree structure for KNN queries
    - Each level has a “cutting dimension”
    - Cycle through the dimensions as you walk down the tree.
  - Complexity: O(NlogN)



Innovation on KNN method– Local Outlier Factor

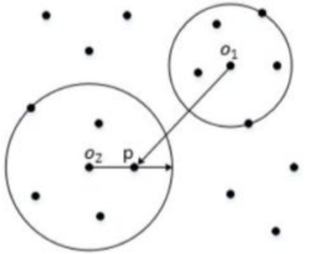
- Backward of KNN Method:
  - Need decide optimal K
  - Distance-based outlier detection models have problems with different densities
- Local Outlier Factor (LOF) [Breunig et al. 2000]
  - Consider relative density
  - Measuring the local deviation of a given data point with respect to its neighbours.



Local Outlier Factor

- Model
  - **K-Distance**
  - **Reachability Distance**:
    - Introduces a smoothing factor:

$$reach-distance_k(p,o) = max\{k-distance(o), d(p,o)\}$$



$$rech-dist_k(p,o_1) = d(p,o_1)$$

$$rech-dist_k(p,o_2) = d_5(o_2)$$

  - **Local Reachability Distance** (lrd) of point p:
    - Inverse of the average reach-dists of the kNNs of p

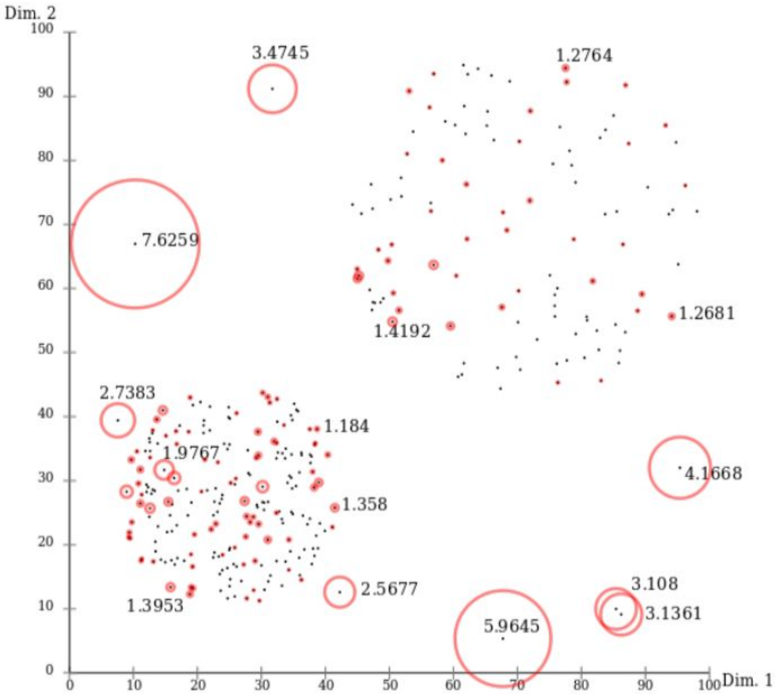
$$lrd_k(p) = 1/\left(\frac{\sum_{o\in N_k(p)} reach-dist_k(p,o)}{|N_k(p)|}\right)$$

Local Outlier Factor

- Model
  - **Local outlier factor (LOF)** of point p
    - Average ratio of lrds of neighbors of p and lrd of p

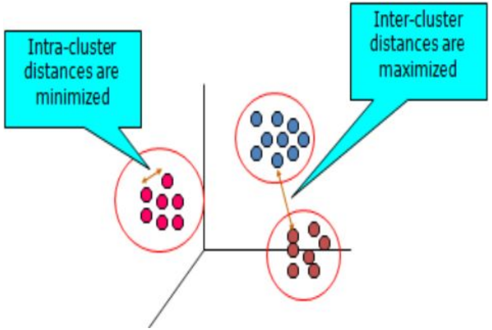
$$LOF_k(p) = \frac{\sum_{o \in N_k(p)} \frac{lrd_k(o)}{lrd_k(p)}}{|N_k(p)|} = \frac{\sum_{o \in N_k(p)} lrd_k(o)}{|N_k(p)|} / lrd_k(p)$$

- Properties
  - LOF ≈ 1: point is in a cluster (region with homogeneous density around the point and its neighbors)
  - LOF >> 1: point is an outlier



Cluster-based Method

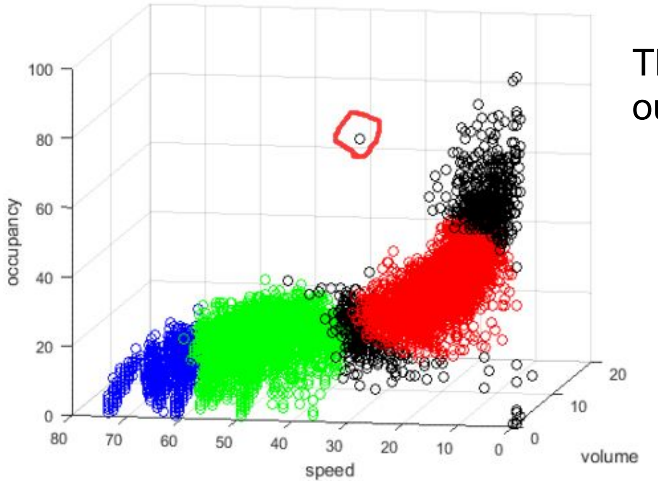
- General Idea
  - Cluster analysis or clustering is the task of assigning a set of object into groups called clusters so that the objects in the same cluster are more similar in some sense to each other than to those in other clusters.



- Implement Distance-based algorithm and Density-based algorithm
  - Distance-based: K-Means
  - Density-based: DBSCAN

K-Means

- Algorithm
    - K-Means clustering aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean.
    - After clustering, an outlier is the point which has a larger distance to the centers of clusters.
- Process of clustering



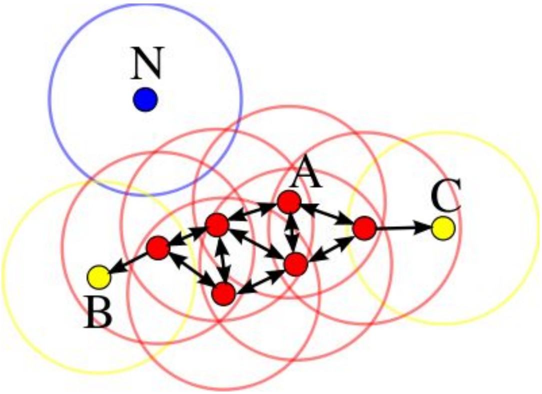
The points that are far from all the centers of clusters are outliers

DBSCAN

- Principle of Density-based spatial clustering of applications with noise
  - Density-based clustering algorithm
  - Based on a set a neighbors of core points to describe tightness.
  - A pair of parameters (ε,MinPts)

DBSCAN

- Algorithm
  - Using (ε,MinPts) to determine the core points: If at least MinPts points are within distance ε.
  - Determine whether a point is reachable or not.
  - All points not reachable from any other points are outliers.
- Process of clustering

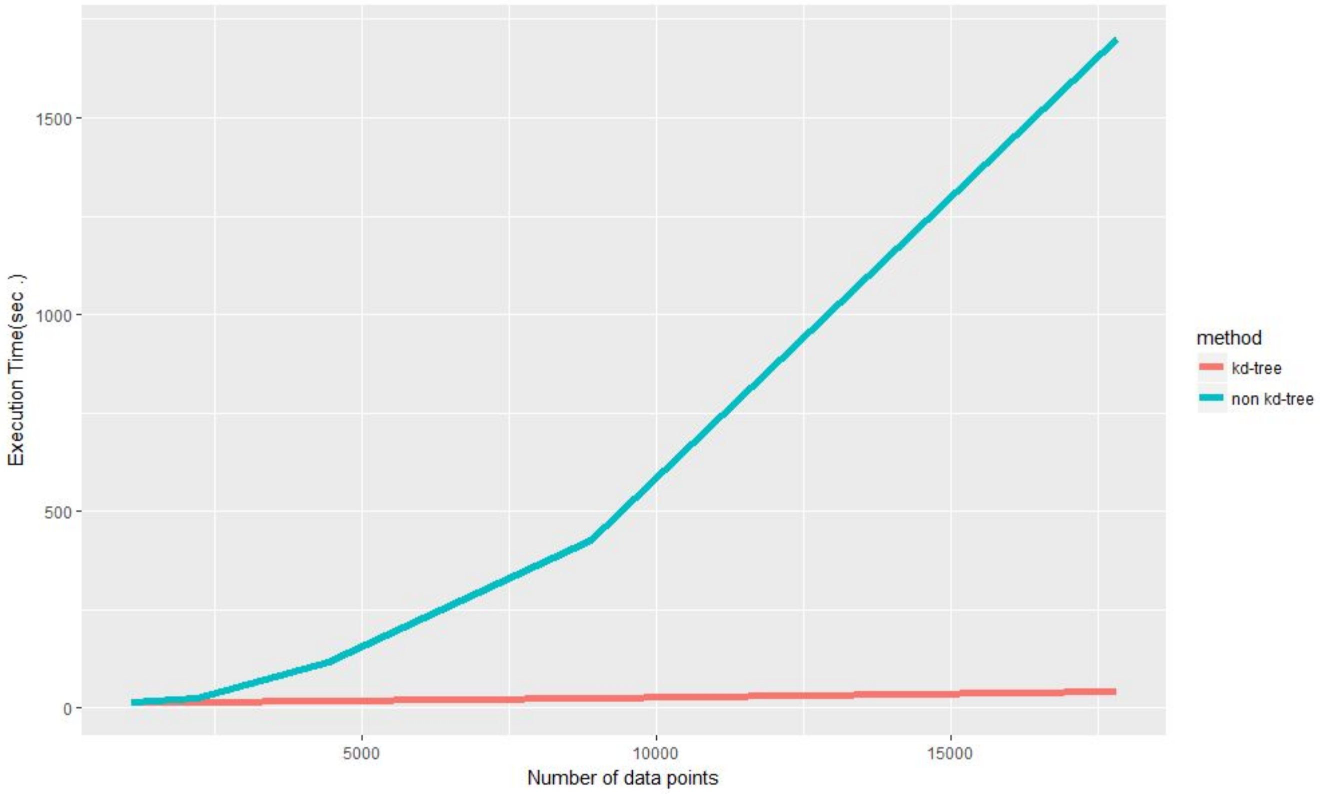


- ❑ We set MinPts = 4. All the red point are core points.
- ❑ Points B and C are not core points
- ❑ Point N is an outlier.

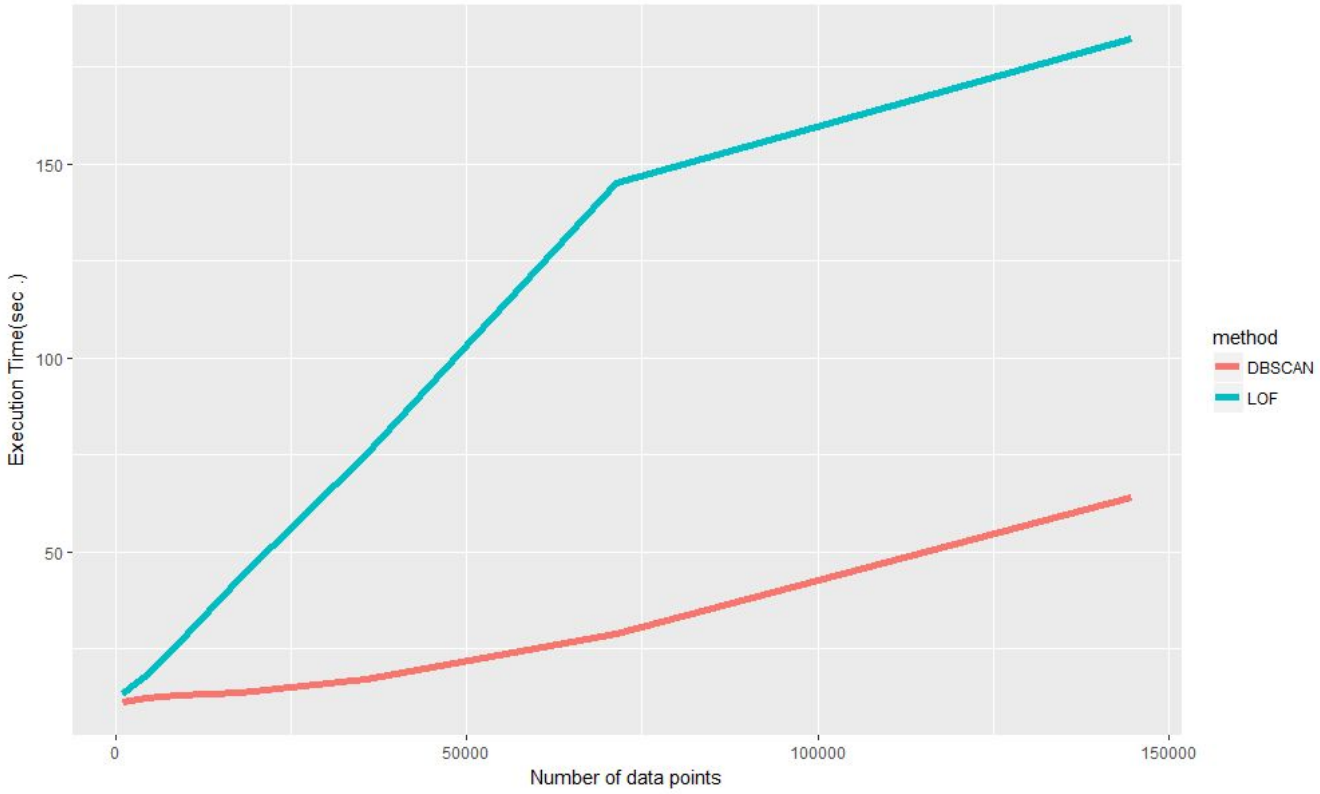
Comparison between K-Means and DBSCAN

- Principle
  - K-Means: Distance-based
  - DBSCAN: Density-based
- Parameters
  - K-Means: Need to define how many clusters in advance.
  - DBSCAN: The number of clusters is determined by the parameter pair: (ε,MinPts).

Performance Result Of Kd-tree & Naïve method



Performance Result Of LOF & DBSCAN



Referenece & Code Repo

[BL94] V. Barnett and T. Lewis. Outliers in Statistical Data. John Wiley and Sons, New York, 1994.
[KN98] Edwin Knorr and Raymond Ng. Algorithms for mining distance-based outliers in large datasets. In Proc. of the VLDB Conference, pages 392-403, New York, USA, September 1998.
[CHT00] Palo Alto, Murray Hill and Taejon. Efficient Algorithms for Mining Outliers from Large Data Sets. ACM Sigmod Record. ACM, 2000, 29(2): 427-438.
[DBSCAN] Density-based spatial clustering of applications with noise (DBSCAN)https://en.wikipedia.org/wiki/DBSCAN
[BKNS00]Breunig, M. M.; Kriegel, H.-P.; Ng, R. T.; Sander, J. (2000). LOF: Identifying Density-based Local Outliers (PDF). Proceedings of the 2000 ACM SIGMOD International Conference on Management ofData. SIGMOD.

Find us on GitHub:

https://github.com/PrinceNathaniel/NYUBigDataProject