LOF: Identifying Density-Based Local Outliers

Wensi DING

July 19, 2016

Outline

Introduction

LOF se base sur l'algorithme de **k plus proches voisins**

Pourquoi LOF ?

- résultat binaire **vs** score
- anomalies globales **vs** anomalies locales

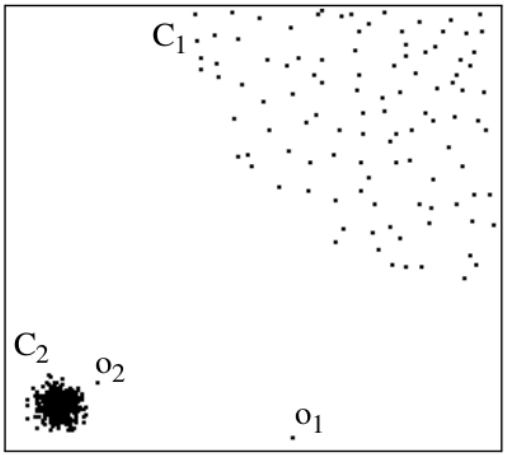

Figure 1: 2-*d* dataset DS1

Quelques définitions

dataset : D
points : o,p,q
cluster : C
distance entre deux points : d(p, q)

k-distance of an object p

k-distance(p) : la distance $d(p, o)$ entre p et un point $o \in D$ tel que

- au moins k points $o\prime \in D\backslash\{p\}$ tel que $d(p, o\prime) \leq d(p, o)$
- au plus k-1 points $o\prime \in D\backslash\{p\}$ tel que $d(p, o\prime) < d(p, o)$

k-distance neighborhood of an object p

$N_{k-distance(p)}(p) = \{q \in D\backslash\{p\} \mid d(p, q) \leq k - distance(p)\}$

Quelques définitions

reachability distance of an object p w.r.t. object o

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

local reachability density of an object p

$$lrd_{MinPts}(p) = 1 / \left(\frac{\sum_{o \in N_{MinPts}(p)} reach\text{-}dist_{MinPts}(p, o)}{|N_{MinPts}(p)|} \right)$$

(local) outlier factor of an object p

$$LOF_{MinPts}(p) = \frac{\sum_{o \in N_{MinPts}(p)} \frac{lrd_{MinPts}(o)}{lrd_{MinPts}(p)}}{|N_{MinPts}(p)|}$$

Quelques propriétés

- $LOF \approx 1$ pour les points au centre du cluster
- LOF ne change pas monotonement avec MinPts

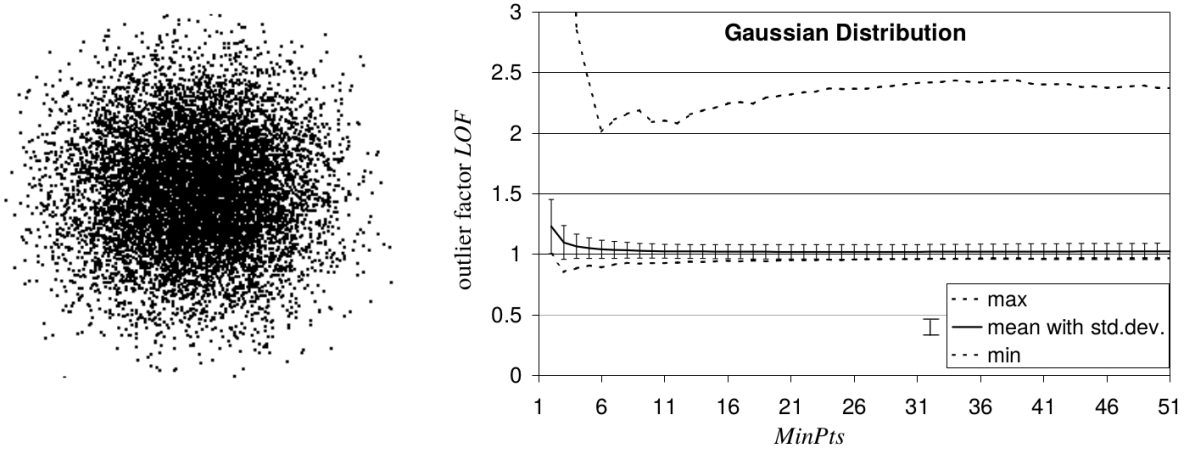

Figure 7: Fluctuation of the outlier-factors within a Gaussian cluster

Choix de MinPts

Méthode heuristique :

$$LOF(p) = max\{LOF_{MinPts}(p) \mid MinPts \in [MinPtsLB, MinPtsUB]\}$$

Guide pour MinPtsLB:

- $MinPtsLB > 10$ pour enlever les fluctuations statistiques
- $MinPtsLB \leq$ minimum du nombre de points dans un cluster normal

Guide pour MinPtsUB:

- $MinPtsUB \geq$ maximum du nombre de points dans un cluster anormal

Complexité: $knn + O(n)$

Exemple

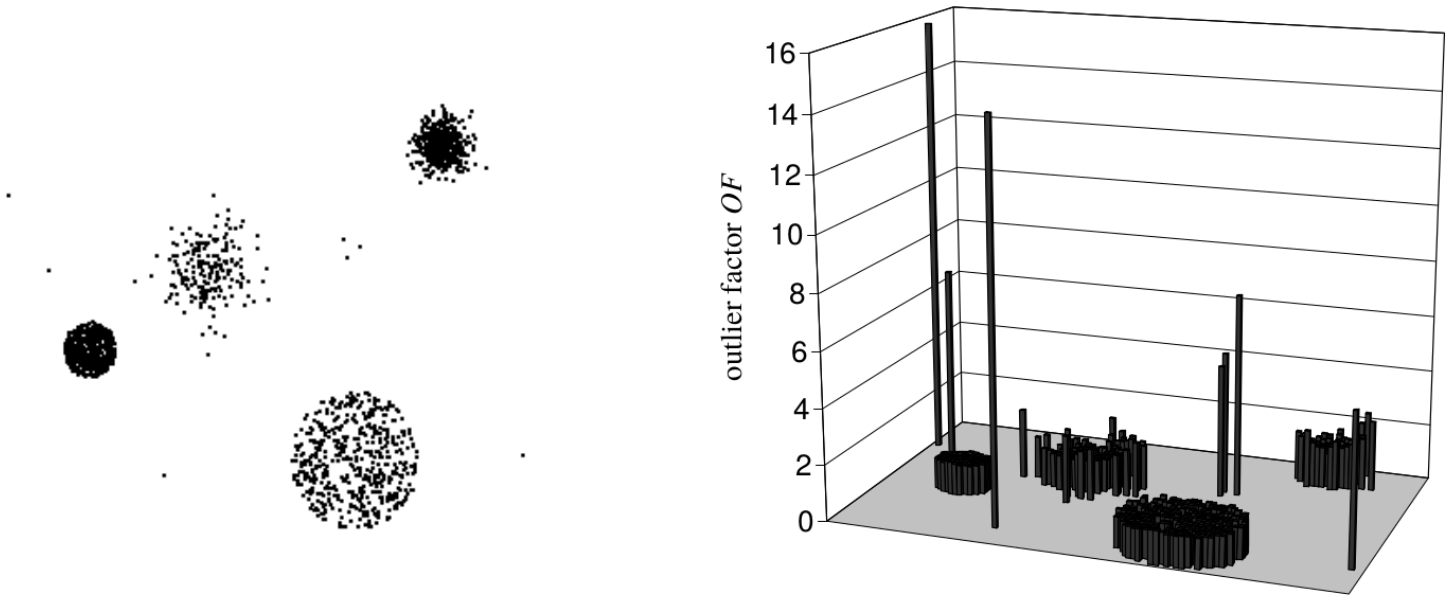

Figure 9: Outlier-factors for points in a sample dataset (*MinPts*=40)

Référence

- Markus M. Breunig, Hans-Peter Kriegel, Raymond T. Ng, Jörg Sander (2000). LOF: Identifying Density-Based Local Outliers. ACM, 1-58113-218-2/00/05