15-826: Multimedia Databases and Data Mining

Outlier detection

C. Faloutsos

Citation:

LOCI: Fast Outlier Detection Using The Local Correlation Integral,

Spiros Papadimitriou,

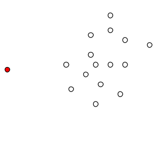
Hiroyuki Kitagawa,

Phillip B. Gibbons,

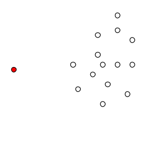
Christos Faloutsos

ICDE 2003, Bangalore, India

What is an outlier?



What is an outlier?



"An outlier is an observation that ***deviates*** so much from other observations as to arouse suspicions that it was generated by a different ***mechanism***" [Hawkins 80]

Why outlier detection?

Why outlier detection?

- Rare events, exceptions, alerts
- Surveillance / auditing
- Health monitoring
- Stock market analysis
- cyber-security
- ...

CMU SCS

Requirements

- Fast
  - Large datasets
  - Single pass over the data
- Multi-dimensional
  - Arbitrary (numeric) tuples
- Limited/simple user intervention
  - No “magic cut-offs”
  - Intuitive criteria for flagging outliers

CMU SCS

Assumptions

- Metric space
  - Distance provided (by “expert”)

OR

- Vector space (*d* dimensions) with some Lp norm

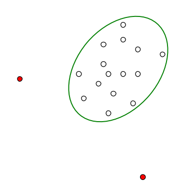
CMU SCS

Overview

- Related work
- Definitions
- Approximate estimation
- Experimental results

CMU SCS

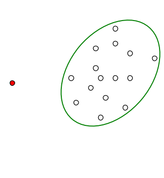
Distribution Based



- Assume data are drawn from a certain distribution (model)
  - E.g., multivariate Gaussian
- Estimate the model
- Identify points that do not fit well

CMU SCS

Distribution Based

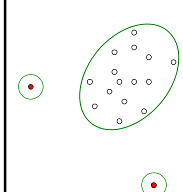


- Assume data are drawn from a certain distribution (model)
  - E.g., multivariate Gaussian
- Estimate the model
- Identify points that do not fit well

- Prior knowledge of distribution, ***or***
- Expensive statistical tests to determine suitable model
- Typically few degrees of freedom
  - E.g., $O(dim^2)$ (covariance matrix)

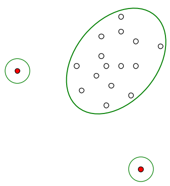
CMU SCS

Clustering by-products



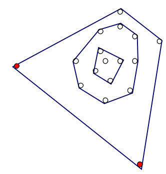
- Outlier is a “cluster”
  - One point
  - Very few points (“micro-cluster”)

Clustering by-products



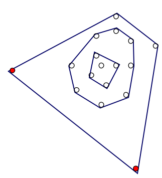
- Outlier is a “cluster”
  - One point
  - Very few points (“micro-cluster”)
- Potentially much unnecessary work
- Implicit outlier-ness criteria

Depth based
[Johnson, Kwok, Ng 98]



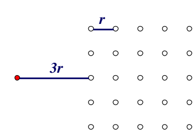
- Computational geometry
- Compute several layers of convex hull — *depth contours* [Tukey 75]
- Points in outer layers are more likely to be outliers

Depth based
[Johnson, Kwok, Ng 98]



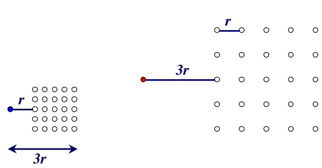
- Computational geometry
- Compute several layers of convex hull — *depth contours* [Tukey 75]
- Points in outer layers are more likely to be outliers
- Exponential growth with dimensionality
  - Unsuitable for high dimensions

Distance based
Overview [Knorr, Ng 98]



- Point x is an outlier iff
  #(points at distance > *R*) > a % #points
  where *R*, *a:* are global parameters

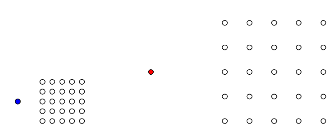
Distance based
Assumptions [Knorr, Ng 98]



- One global criterion
- Local density variations

Global may not be enough...

Density based (LOF)
Overview [Breuning, Kriegel, Ng etal 00]

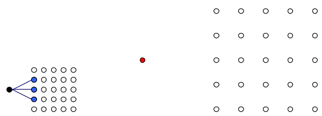


- Compute (smoothed) average neighbor distance (over *m* neighbors)
- Compare against average of averages

CMU SCS

Density based (LOF)

Overview [Breuning, Kriegel, Ng etal 00]



- Local Outlier Factor (LOF)

$$\mathrm{LOF}(p, m) = \frac{\text{avg. distance of } m \text{ nearest neighbours of } p}{\text{avg. of avg. distance of m nearest neighbours over the } m \text{ nearest neighbours of p}}$$

CMU SCS

Density based

Assumptions

- Multiple granularity
- Here, $m > 9$ (micro-cluster size)
- If $m < 9$,then none are outliers (LOF = 1)

Local may not be enough either…

CMU SCS

Contributions

- Multi-scale
  - Complete distance distribution…
  - …around each point
- Fast estimation
  - Practically linear in #points and dimensions
  - First use of approximate estimation in outlier detection
- “Natural” cut-off
  - Probabilistically intuitive, data-dictated

CMU SCS

Overview

- Related work
- Definitions
- Approximate estimation
- Experimental results

CMU SCS

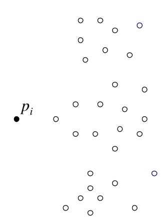
Main questions/ideas:

- Q1: What is a neighbor?
- Q2: how to measure difference from neighbors?
- Q3: how much difference is ‘too much’?

CMU SCS

Comparing distances

Q1: who is a neighbor?

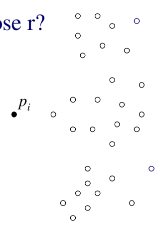


CMU SCS

Comparing distances

Q1: who is a neighbor? A: all within distance r

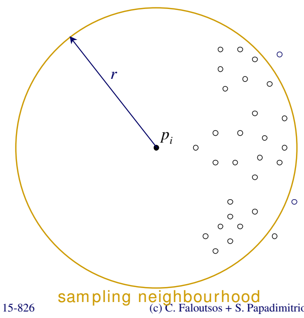
Q1': how to choose r?



15-826 (c) C. Faloutsos + S. Papadimitriou (2005) 25

CMU SCS

Comparing distances



15-826 (c) C. Faloutsos + S. Papadimitriou (2005) 26

CMU SCS

Main questions/ideas:

- Q1: What is a neighbor?
- Q2: how to measure difference from neighbors?
- Q3: how much difference is 'too much'?

15-826 (c) C. Faloutsos + S. Papadimitriou (2005) 27

CMU SCS

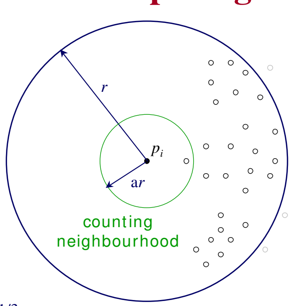
Main questions/ideas:

- Q1: What is a neighbor?
- Q2: how to measure difference from neighbors?
- A: count the number of neighbors!

15-826 (c) C. Faloutsos + S. Papadimitriou (2005) 28

CMU SCS

Comparing distances



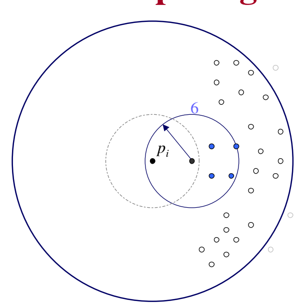
$n(p_i, r) = 2$

a = 1/3

15-826 (c) C. Faloutsos + S. Papadimitriou (2005) 29

CMU SCS

Comparing distances



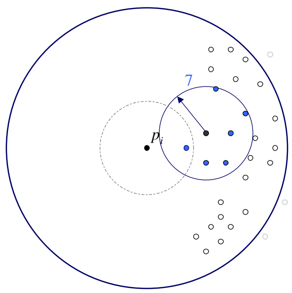
$n(p_i, r) = 2$

$\hat{n}(p_i, r, a) = \frac{2+6}{19}$

15-826 (c) C. Faloutsos + S. Papadimitriou (2005) 30

CMU SCS

Comparing distances

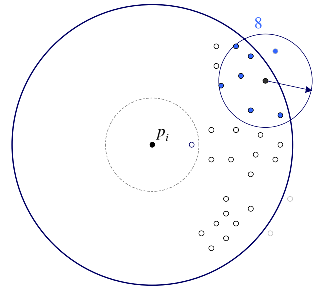


$n(p_i, r) = 2$

$\bar{n}(p_i, r, \alpha) = \frac{2+6+7}{19}$

CMU SCS

Comparing distances



$n(p_i, r) = 2$

$\bar{n}(p_i, r, \alpha) = \frac{2+6+7+8+\cdots}{19}$

CMU SCS

Main questions/ideas:

- Q1: What is a neighbor?
- Q2: how to measure difference from neighbors?
- Q3: how much difference is 'too much'?

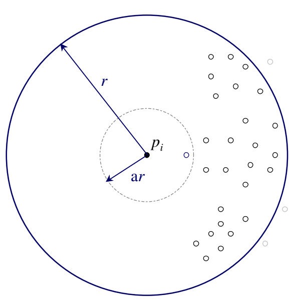
CMU SCS

Main questions/ideas:

- Q1: What is a neighbor?
- Q2: how to measure difference from neighbors?
- Q3: how much difference is 'too much'?
- A: 3 std away (or *k* std)

CMU SCS

Comparing distances



$n(p_i, r) = 2$

$\bar{n}(p_i, r, \alpha) = 8.21$ (average)

$\sigma_n(p_i, r, \alpha) = 2.01$ (std. dev.)

$\bar{n}(p_i, r, \alpha) - n(p_i, \alpha r) > 3\sigma_n(p_i, r, \alpha)$

k

Neighbor counts:
2 6 6 6 6 7 7 7 7 8
8 8 8 8 8 9 9 9 9 9 9
10 10 10 10 10 12 12

CMU SCS

Main questions/ideas:

- Q1: What is a neighbor?
- A: everybody within distance *r*
  - Q1': how to choose *r* ?
  - A:
- Q2: how to measure difference from neighbors?
- Q3: how much difference is 'too much'?

CMU SCS

Main questions/ideas:

- Q1: What is a neighbor?
- A: everybody within distance *r*
  - Q1’: how to choose *r* ?
  - A: don’t choose it – MANY *r* ‘s!
- Q2: how to measure difference from neighbors?
- Q3: how much difference is ‘too much’?

15-826 (c) C. Faloutsos + S. Papadimitriou (2005) 37

CMU SCS

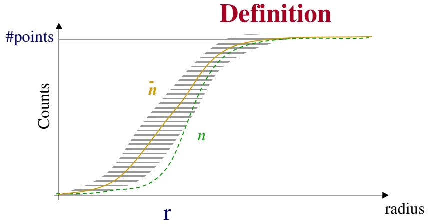
End result:

- LOCI plot for each point *p*
- (LOCI = LOcal Correlation Integral)

15-826 (c) C. Faloutsos + S. Papadimitriou (2005) 38

CMU SCS

LOCI plot
Definition



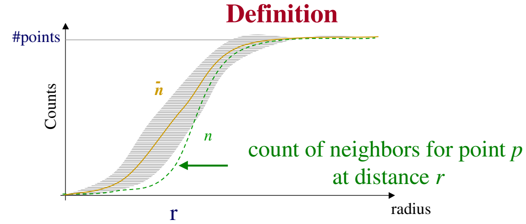
- Plot of

$$\left.\begin{matrix}\langle \text{count } n \rangle \\ \langle \text{avg } \hat{n} \rangle \pm k_\sigma \times \langle \text{std dev } \hat{\sigma}_n \rangle \end{matrix}\right\} \text{vs. radius}$$

15-826 (c) C. Faloutsos + S. Papadimitriou (2005) 39

CMU SCS

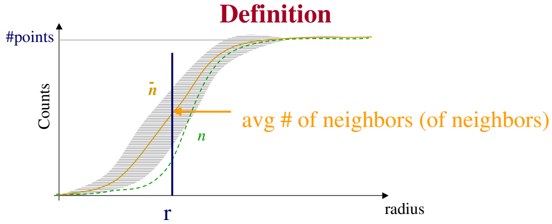
LOCI plot
Definition



15-826 (c) C. Faloutsos + S. Papadimitriou (2005) 40

CMU SCS

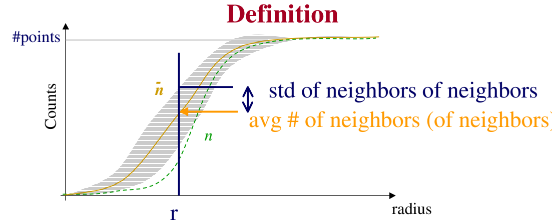
LOCI plot
Definition



15-826 (c) C. Faloutsos + S. Papadimitriou (2005) 41

CMU SCS

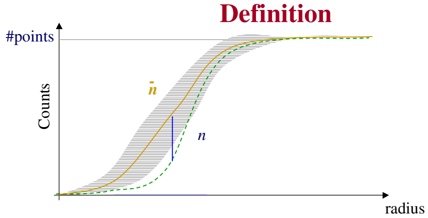
LOCI plot
Definition



15-826 (c) C. Faloutsos + S. Papadimitriou (2005) 42

CMU SCS

LOCI plot
Definition

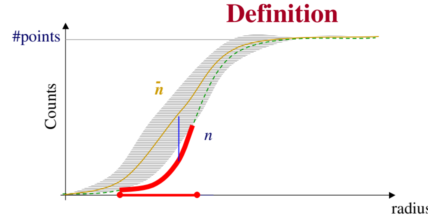


- Point is an outlier (at "scale" r) iff the green line is outside the gray band!

15-826 (c) C. Faloutsos + S. Papadimitriou (2005) 43

CMU SCS

LOCI plot
Definition

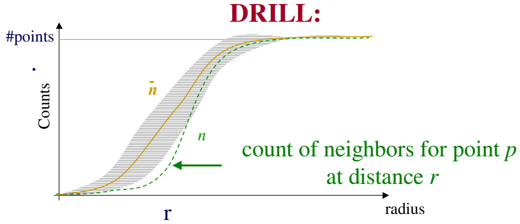


- Point is an outlier (at "scale" r) iff the green line is outside the gray band!

15-826 (c) C. Faloutsos + S. Papadimitriou (2005) 44

CMU SCS

LOCI plot
DRILL:

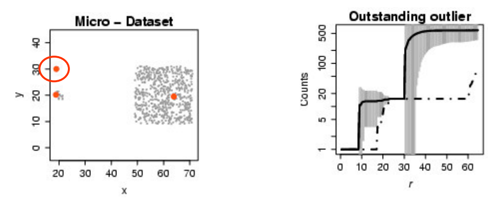


Suppose that the green curve is a line in log-log scales with slope 1.58. What is your conclusion about point p ?

15-826 (c) C. Faloutsos + S. Papadimitriou (2005) 45

CMU SCS

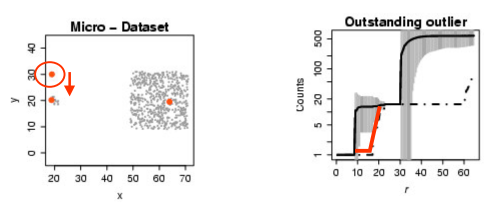
LOCI plot
Properties



15-826 (c) C. Faloutsos + S. Papadimitriou (2005) 46

CMU SCS

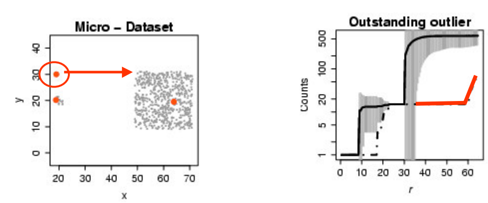
LOCI plot
Properties



15-826 (c) C. Faloutsos + S. Papadimitriou (2005) 47

CMU SCS

LOCI plot
Properties



15-826 (c) C. Faloutsos + S. Papadimitriou (2005) 48

CMU SCS

Overview

- Related work
- Definitions
- Approximate estimation
- Experimental results

CMU SCS

Speed?

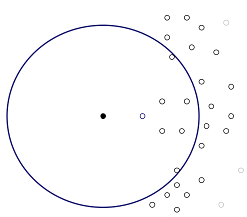
- seems quadratic or worse:
  - for every radius, count the neighbors,
  - the neighbors of the neighbors
  - and the std of them
- How to 'cut corners'?

CMU SCS

Box counting

Introduction by example

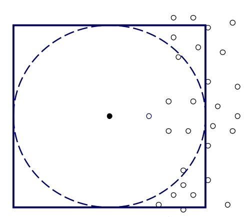
Exact



CMU SCS

Box counting

Introduction by example

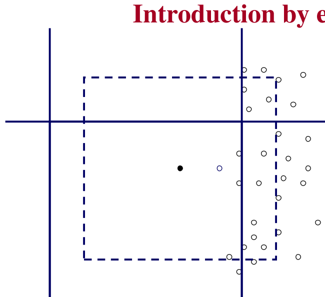


Approximate:
impose a grid of side *r*
- squares (L-infinity)

CMU SCS

Box counting

Introduction by example



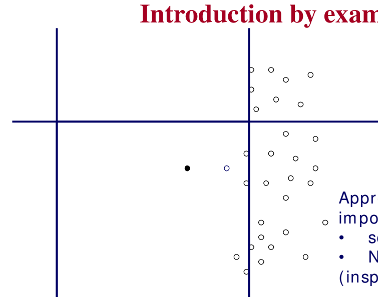
Approximate:
impose a grid of side *r*
- squares (L-infinity)
- Not necessarily centered

(inspired by box-counting)

CMU SCS

Box counting

Introduction by example



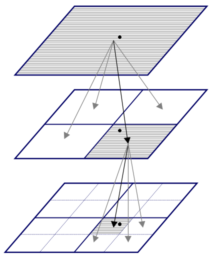
Approximate:
impose a grid of side *r*
- squares (L-infinity)
- Not necessarily centered

(inspired by box-counting)

CMU SCS

SKIP

Quad-trees



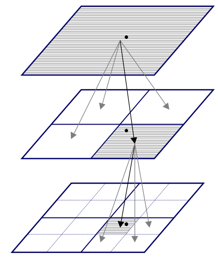
- Cell side (radius) can be arbitrary, however…
- Quad-trees:
  - Split each cell into 2^d sub-cells
  - Small depth, easy updates

Radii = powers of 2

CMU SCS

SKIP

Quad-trees



- Store only counters

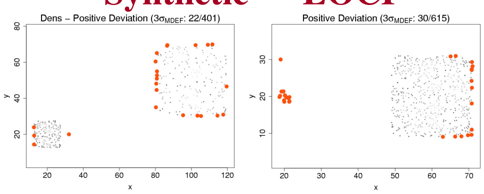
CMU SCS

Overview

- Related work
- Definitions
- Approximate estimation
- Experimental results

CMU SCS

Synthetic — LOCI

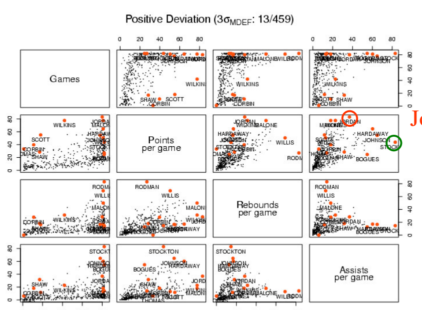


- Captures outstanding outlier

- Captures outstanding outlier…
- …and entire micro-cluster

CMU SCS

NBA — LOCI



Jordan

Stockton

CMU SCS

Conclusion

- Multi-scale
  - Complete distance distribution…
  - …around each point
- Fast estimation
  - Box-counting, practically linear
  - First use of approximate estimation in outlier detection
- "Natural" cut-off
  - Probabilistically intuitive, data-dictated

CMU SCS

References

[Knorr, Ng 98]

Algorithms for mining distance-based outliers in large datasets, Proc. VLDB, 392-403, 1998

[Breunig, Kriegel, Ng, etal 00]

LOF: Identifying density-based local outliers, Proc. SIGMOD, 93-104, 2000

[Johnson, Kwok, Ng 98]

Fast computation of 2-dimensional depth contours, Proc. KDD, 224-228, 1998

[Hawkins 80]

Identification of Outliers, Chapman and Hall, London, 1980