6.6 Density-based Approaches

General idea

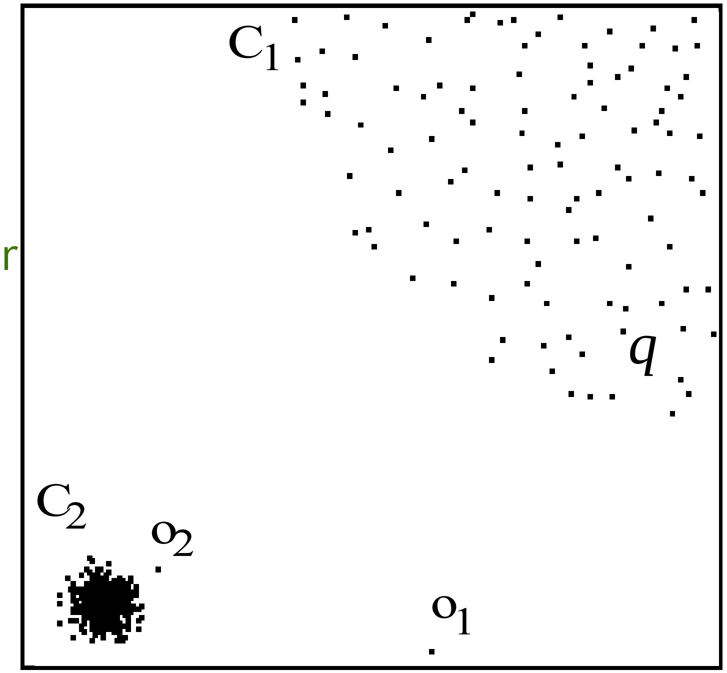
- Compare the density around a point with the density around its local neighbors
- The relative density of a point compared to its neighbors is computed as an outlier score
- Approaches also differ in how to estimate density

Basic assumption

- The density around a normal data object is similar to the density around its neighbors
- The density around an outlier is considerably different to the density around its neighbors

Local Outlier Factor (LOF) [Breunig et al. 1999], [Breunig et al. 2000]

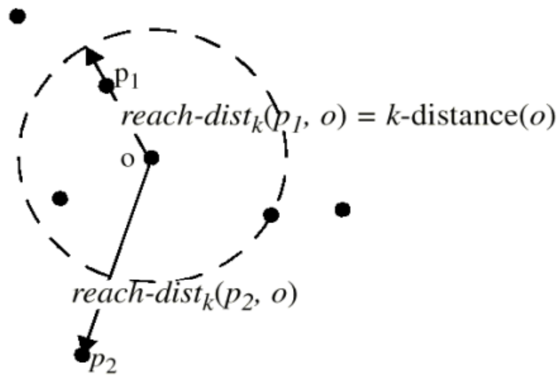
- Motivation:
  - Distance-based outlier detection models have problems with different densities
  - How to compare the neighborhood of points from areas of different densities?
  - Example
    - DB(ε,π)-outlier model
      - » Parameters ε and π cannot be chosen so that o_2 is an outlier but none of the points in cluster C_1 (e.g. *q*) is an outlier
    - Outliers based on kNN-distance
      - » kNN-distances of objects in C_1 (e.g. *q*) are larger than the kNN-distance of o_2



- Solution: consider relative density

– Model
  - Reachability distance
    – Introduces a smoothing factor

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



  - Local reachability distance (lrd) of point *p*
    – Inverse of the average reach-dists of the *k*NNs of *p*

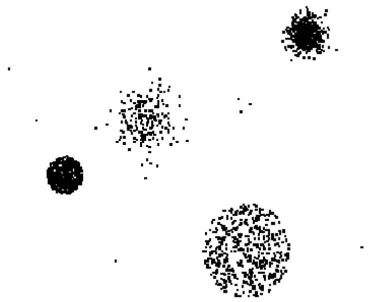
$$lrd_k(p) = 1 \Big/ \left(\frac{\sum_{o \in kNN(p)} reach\text{-}dist_k(p,o)}{Card\left(kNN(p)\right)} \right)$$

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

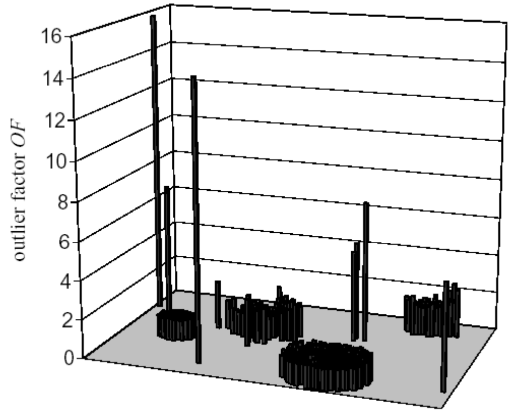
$$LOF_k(p) = \frac{\sum_{o \in kNN(p)} \frac{lrd_k(o)}{lrd_k(p)}}{Card\left(kNN(p)\right)}$$

6.6 Density-based Approaches

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



Data set



LOFs (*MinPts* = 40)

– Discussion
  - Choice of *k* (*MinPts* in the original paper) specifies the reference set
  - Originally implements a local approach (resolution depends on the user’s choice for *k*)
  - Outputs a scoring (assigns an LOF value to each point)

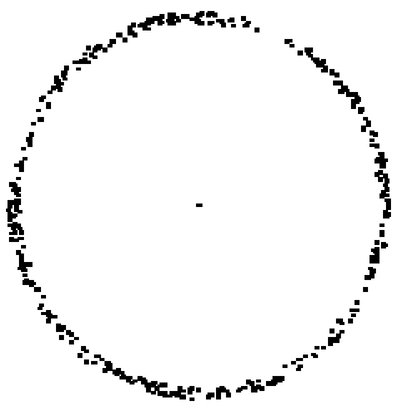
Variants of LOF

- Mining top-*n* local outliers [Jin et al. 2001]
  - Idea:
    - Usually, a user is only interested in the top-*n* outliers
    - Do not compute the LOF for all data objects => save runtime
  - Method
    - Compress data points into micro clusters using the CFs of BIRCH [Zhang et al. 1996]
    - Derive upper and lower bounds of the reachability distances, lrd-values, and LOF-values for points within a micro clusters
    - Compute upper and lower bounds of LOF values for micro clusters and sort results w.r.t. ascending lower bound
    - Prune micro clusters that cannot accommodate points among the top-*n* outliers (*n* highest LOF values)
    - Iteratively refine remaining micro clusters and prune points accordingly

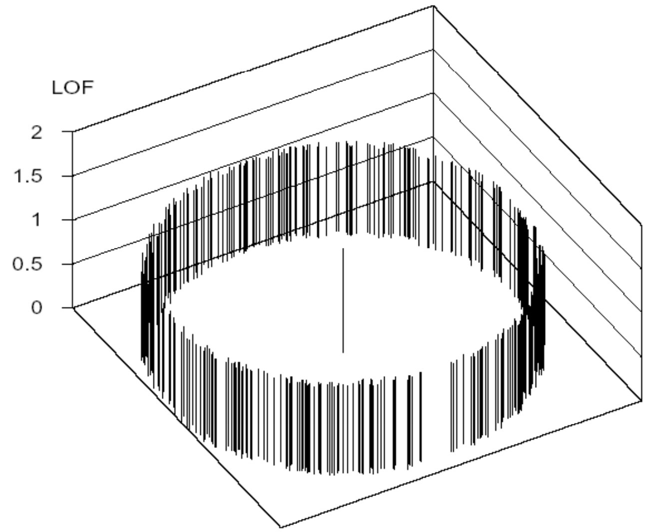
Variants of LOF (cont.)

– Connectivity-based outlier factor (COF) [Tang et al. 2002]

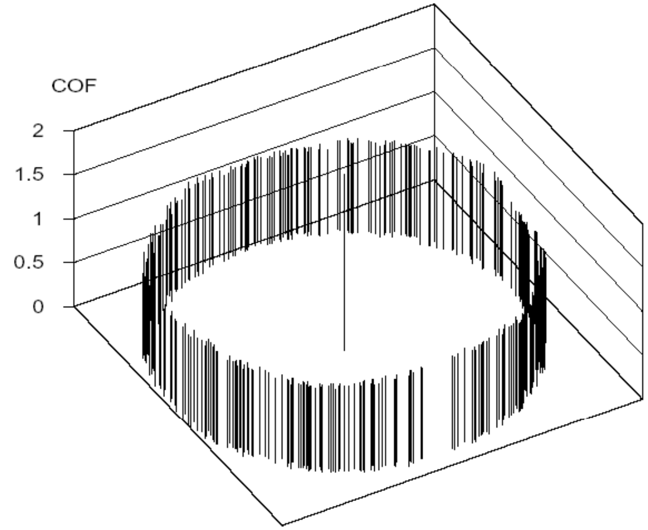
- Motivation
  - In regions of low density, it may be hard to detect outliers
  - Choose a low value for *k* is often not appropriate
- Solution
  - Treat "low density" and "isolation" differently
- Example



Data set



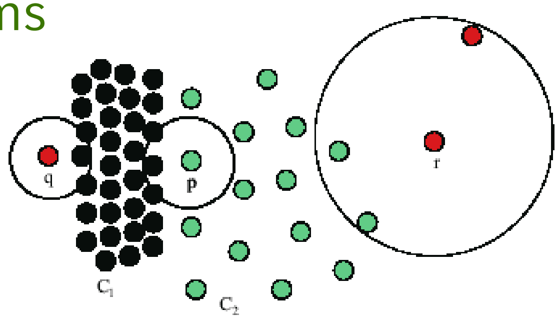
LOF



COF

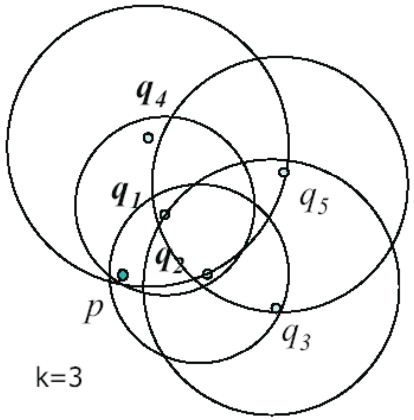
Influenced Outlierness (INFLO) [Jin et al. 2006]

– Motivation
  - If clusters of different densities are not clearly separated, LOF will have problems



Point *p* will have a higher LOF than points *q* or *r* which is counter intuitive

– Idea
  - Take symmetric neighborhood relationship into account
  - Influence space (kIS(p)) of a point p includes its kNNs (kNN(p)) and its reverse kNNs (RkNN(p))



$$kIS(p) = kNN(p) \cup RkNN(p))$$
$$= \{q_1, q_2, q_4\}$$

- Model
  - Density is simply measured by the inverse of the *k*NN distance, i.e.,

    $den(p) = 1/k\text{-distance}(p)$

  - Influenced outlierness of a point p

$$INFLO_k(p) = \frac{\sum_{o \in kIS(p)} den(o) \Big/ Card(kIS(p))}{den(p)}$$

  - INFLO takes the ratio of the average density of objects in the neighborhood of a point *p* (i.e., in *k*NN(*p*) ∪ R*k*NN(*p*)) to *p*'s density

- Proposed algorithms for mining top-*n* outliers
  - Index-based
  - Two-way approach
  - Micro cluster based approach

- Properties
  - Similar to LOF
  - INFLO ≈ 1: point is in a cluster
  - INFLO >> 1: point is an outlier
- Discussion
  - Outputs an outlier score
  - Originally proposed as a local approach (resolution of the reference set kIS can be adjusted by the user setting parameter k)

Local outlier correlation integral (LOCI) [Papadimitriou et al. 2003]

- Idea is similar to LOF and variants
- Differences to LOF
  - Take the ε-neighborhood instead of *k*NNs as reference set
  - Test multiple resolutions (here called "granularities") of the reference set to get rid of any input parameter
- Model
  - ε-neighborhood of a point p: N(*p*,ε) = {*q* | *dist*(*p*,*q*) ≤ ε}
  - Local density of an object p: number of objects in N(*p*,ε)
  - Average density of the neighborhood

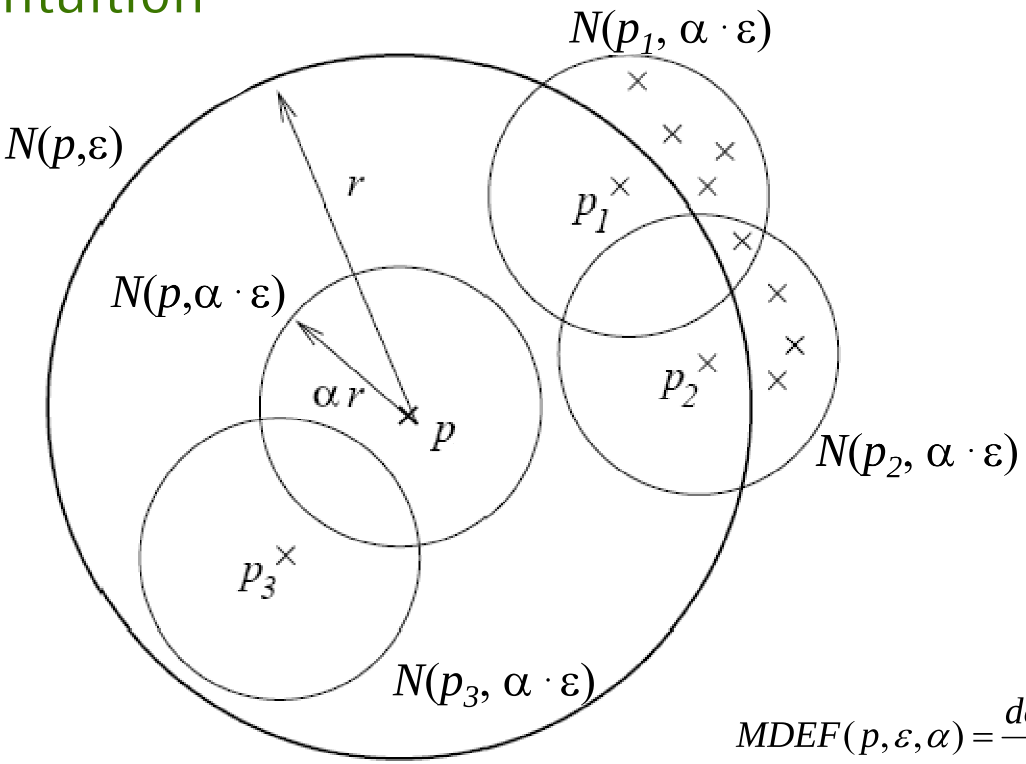
$$den(p,\varepsilon,\alpha) = \frac{\sum_{q \in N(p,\varepsilon)} Card(N(q,\alpha \cdot \varepsilon))}{Card(N(p,\varepsilon))}$$

  - Multi-granularity Deviation Factor (MDEF)

$$MDEF(p,\varepsilon,\alpha) = \frac{den(p,\varepsilon,\alpha) - Card(N(p,\alpha \cdot \varepsilon))}{den(p,\varepsilon,\alpha)} = 1 - \frac{Card(N(p,\alpha \cdot \varepsilon))}{den(p,\varepsilon,\alpha)}$$

6.6 Density-based Approaches

– Intuition



$$den(p,\varepsilon,\alpha) = \frac{\sum_{q \in N(p,\varepsilon)} Card(N(q,\alpha \cdot \varepsilon))}{Card(N(p,\varepsilon))}$$

$$MDEF(p,\varepsilon,\alpha) = \frac{den(p,\varepsilon,\alpha) - Card(N(p,\alpha \cdot \varepsilon))}{den(p,\varepsilon,\alpha)} = 1 - \frac{Card(N(p,\alpha \cdot \varepsilon))}{den(p,\varepsilon,\alpha)}$$

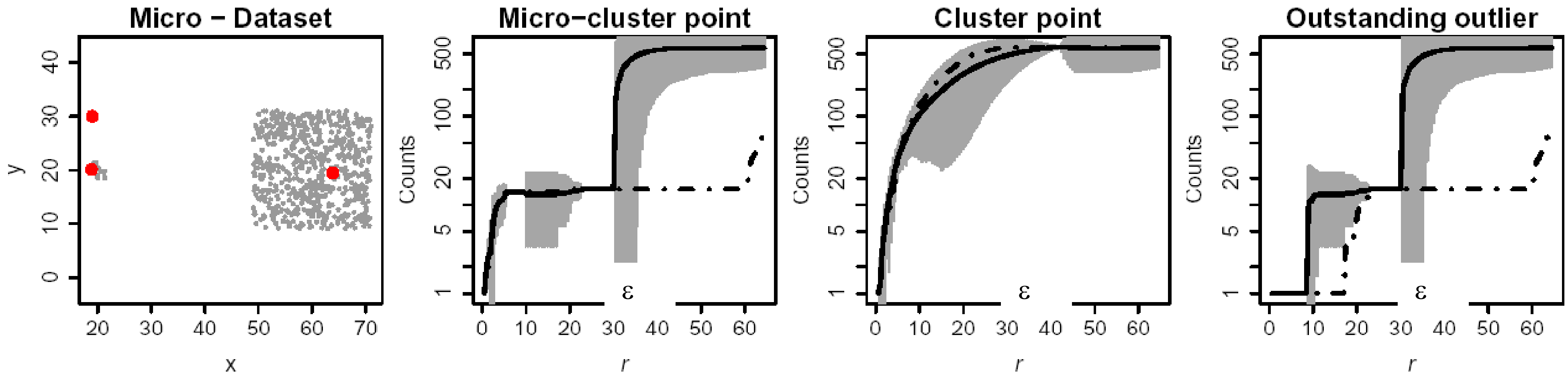
– σMDEF(p,ε,α) is the normalized standard deviation of the densities of all points from $N(p,\varepsilon)$

– Properties

- MDEF = 0 for points within a cluster
- MDEF > 0 for outliers or MDEF > 3·σMDEF => outlier

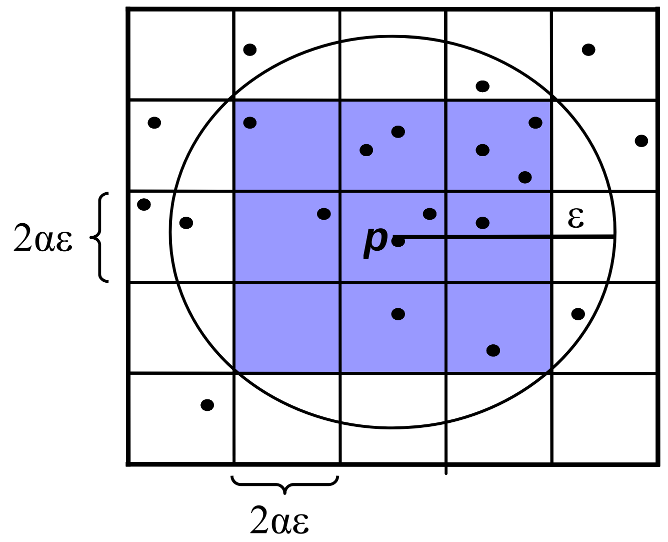
6.6 Density-based Approaches

– Features
  - Parameters ε and α are automatically determined
  - In fact, all possible values for ε are tested
  - LOCI plot displays for a given point *p* the following values w.r.t. ε
    – *Card*(*N*(*p*, α·ε))
    – *den*(*p*, ε, α) with a border of ± 3·σ*den*(*p*, ε, α)



– Algorithms

- Exact solution is rather expensive (compute MDEF values for all possible ε values)
- aLOCI: fast, approximate solution
  - Discretize data space using a grid with side length 2αε
  - Approximate range queries trough grid cells
  - ε - neighborhood of point p: ζ(p,ε) all cells that are completely covered by ε-sphere around *p*
  - Then,



$$Card(N(q,\alpha\cdot\varepsilon))=\frac{\sum_{c_j\in\zeta(p,\varepsilon)} c_j^{\,2}}{\sum_{c_j\in\zeta(p,\varepsilon)} c_j}$$

where c_j is the object count the corresponding cell

  - Since different ε values are needed, different grids are constructed with varying resolution
  - These different grids can be managed efficiently using a Quad-tree

- Discussion
  - Exponential runtime w.r.t. data dimensionality
  - Output:
    - Label: if MDEF of a point > 3·σMDEF then this point is marked as outlier
    - LOCI plot
      - » At which resolution is a point an outlier (if any)
      - » Additional information such as diameter of clusters, distances to clusters, etc.
  - All interesting resolutions, i.e., possible values for ε, (from local to global) are tested

Übersicht

6.2–6.4: Statistisches Modell

6.5–6.6: Modellierung durch räumliche Nähe

6.7: Anpassung verschiedener Modelle an spezielles Problem

6.7 High-dimensional Approaches

- One sample class of adaptations of existing models to a specific problem (high dimensional data)
- Why is that problem important?
  - Some (ten) years ago:
    - Data recording was expansive
    - Variables (attributes) where carefully evaluated whether or not they are relevant for the analysis task
    - Data sets usually contain only a few number of relevant dimensions
  - Nowadays:
    - Data recording is easy and cheap
    - "Everyone measures everything", attributes are not evaluated just measured
    - Data sets usually contain a large number of features
      - » Molecular biology: gene expression data with >1,000 of genes per patient
      - » Customer recommendation: ratings of 10-100 of products per person
      - » …

High-dimensional Approaches

Challenges

- Curse of dimensionality
  - Relative contrast between distances decreases with increasing dimensionality
  - Data are very sparse, almost all points are outliers
  - Concept of neighborhood becomes meaningless

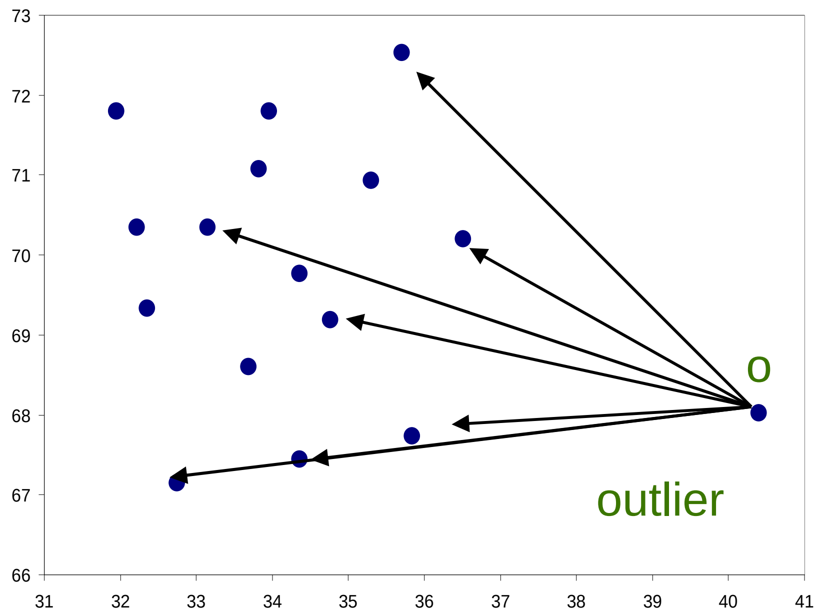
- Solutions
  - Use more robust distance functions and find full-dimensional outliers
  - Find outliers in projections (subspaces) of the original feature space

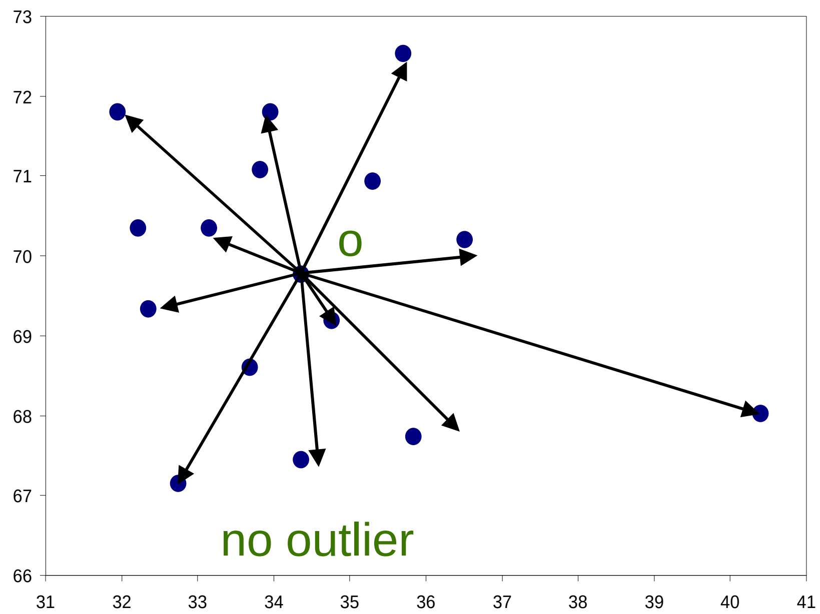
High-dimensional Approaches

ABOD – angle-based outlier degree [Kriegel et al. 2008]

– Rational

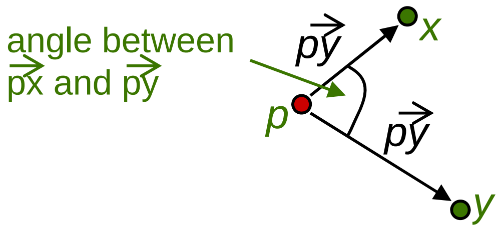
- Angles are more stable than distances in high dimensional spaces (cf. e.g. the popularity of cosine-based similarity measures for text data)
- Object o is an outlier if most other objects are located in similar directions
- Object o is no outlier if many other objects are located in varying directions

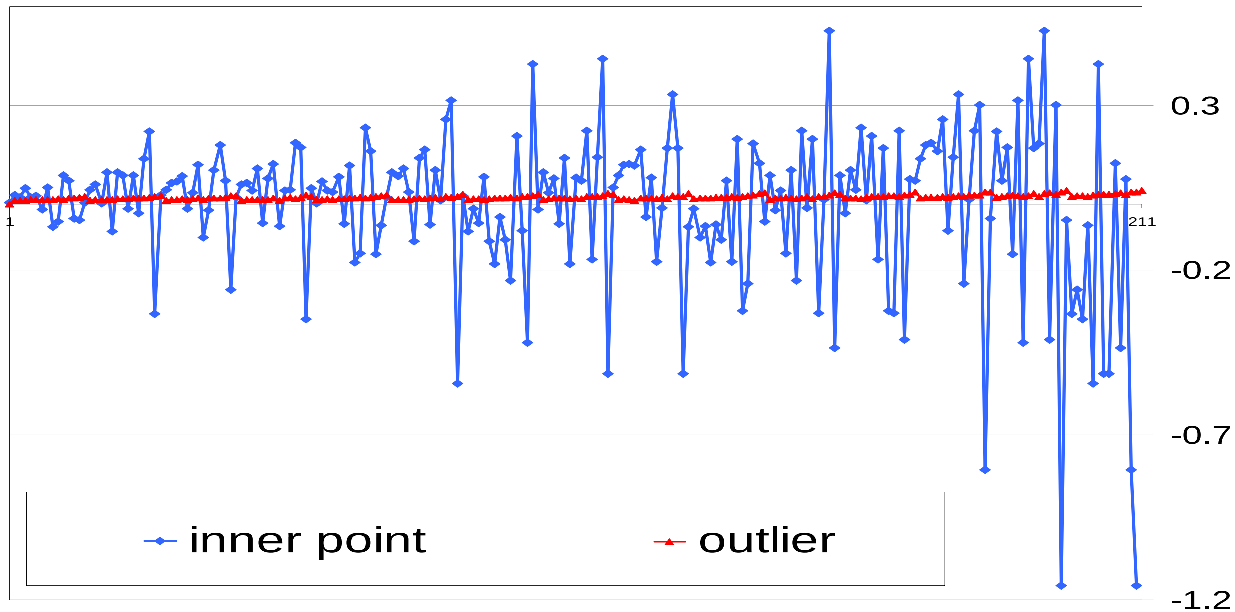




High-dimensional Approaches

- Basic assumption
  - Outliers are at the border of the data distribution
  - Normal points are in the center of the data distribution
- Model
  - Consider for a given point *p* the angle between $\overrightarrow{px}$ and $\overrightarrow{py}$ for any two *x*,*y* from the database
  - Consider the spectrum of all these angles
  - The broadness of this spectrum is a score for the outlierness of a point





High-dimensional Approaches

- Model (cont.)
  - Measure the variance of the angle spectrum
  - Weighted by the corresponding distances (for lower dimensional data sets where angles are less reliable)

$$ABOD(p) = \underset{x,y \in DB}{VAR}\left(\frac{\left\langle \vec{xp}, \vec{yp} \right\rangle}{\left\| \vec{xp} \right\|^2 \cdot \left\| \vec{yp} \right\|^2} \right)$$

  - Properties
    - Small ABOD => outlier
    - High ABOD => no outlier

High-dimensional Approaches

- Algorithms
  - Naïve algorithm is in $O(n^3)$
  - Approximate algorithm based on random sampling for mining top-*n* outliers
    - Do not consider all pairs of other points *x*,*y* in the database to compute the angles
    - Compute ABOD based on samples => lower bound of the real ABOD
    - Filter out points that have a high lower bound
    - Refine (compute the exact ABOD value) only for a small number of points
- Discussion
  - Global approach to outlier detection
  - Outputs an outlier score (inversely scaled: high ABOD => inlier, low ABOD => outlier)

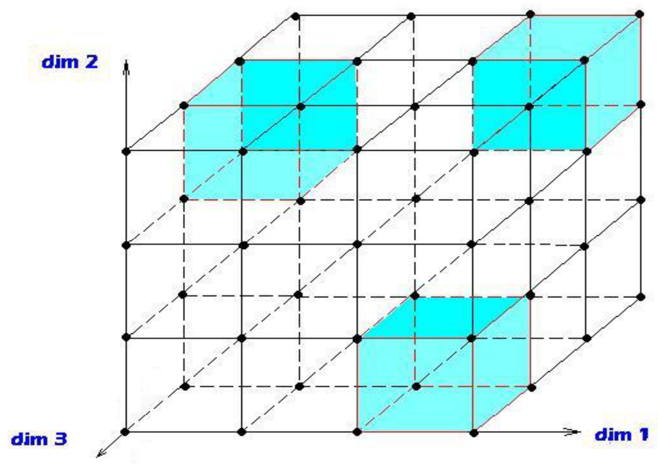
Grid-based subspace outlier detection [Aggarwal and Yu 2000]

– Model
  - Partition data space by an equi-depth grid (Φ = number of cells in each dimension)
  - Sparsity coefficient *S(C)* for a *k*-dimensional grid cell *C*

$$S(C) = \frac{count(C) - n \cdot (1/\Phi)^k}{\sqrt{n \cdot (1/\Phi)^k \cdot (1 - (1/\Phi)^k)}}$$

where *count*(*C*) is the number of data objects in C

  - *S(C)* < 0 => *count*(*C*) is lower than expected
  - Outliers are those objects that are located in lower-dimensional cells with negative sparsity coefficient



$\Phi = 3$

High-dimensional Approaches

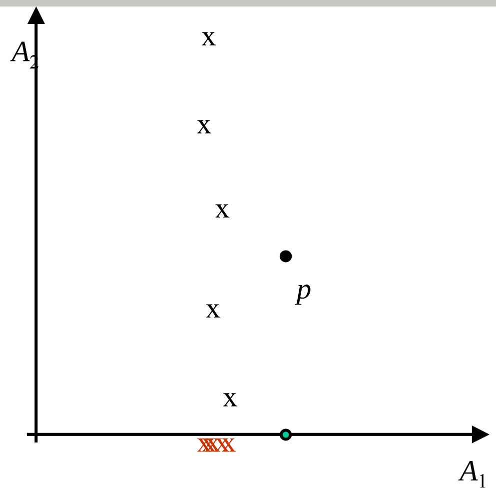
- Algorithm
  - Find the *m* grid cells (projections) with the lowest sparsity coefficients
  - Brute-force algorithm is in $O(\Phi^d)$
  - Evolutionary algorithm (input: *m* and the dimensionality of the cells)

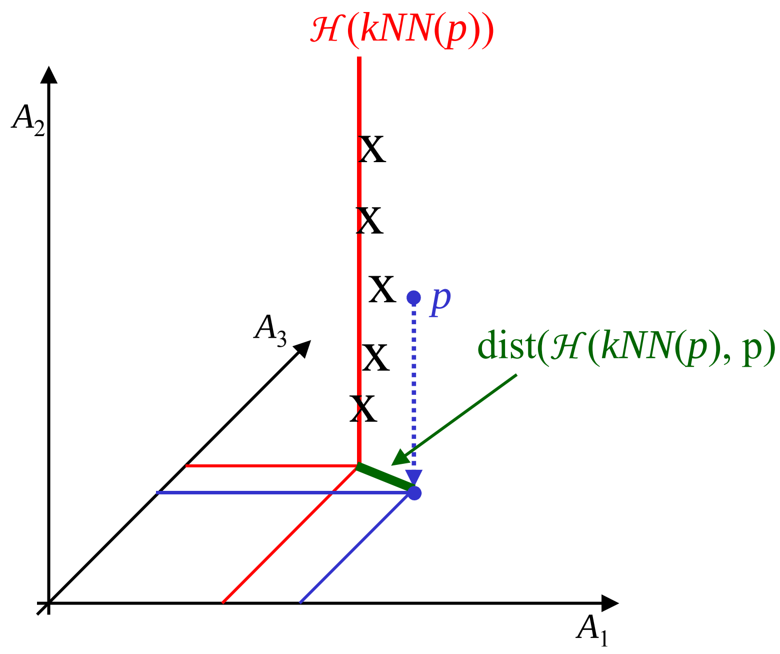
- Discussion
  - Results need not be the points from the optimal cells
  - Very coarse model (all objects that are in cell with less points than to be expected)
  - Quality depends on grid resolution and grid position
  - Outputs a labeling
  - Implements a global approach (key criterion: globally expected number of points within a cell)

High-dimensional Approaches

SOD – subspace outlier degree [Kriegel et al. 2009]

- Motivation
  - Outliers may be visible only in subspaces of the original data
- Model
  - Compute the subspace in which the *k*NNs of a point *p* minimize the variance
  - Compute the hyperplane $\mathcal{H}(kNN(p))$ that is orthogonal to that subspace
  - Take the distance of *p* to the hyperplane as measure for its "outlierness"





High-dimensional Approaches

– Discussion
  - Assumes that *k*NNs of outliers have a lower-dimensional projection with small variance
  - Resolution is local (can be adjusted by the user via the parameter *k*)
  - Output is a scoring (SOD value)

Outline

Summary

- Historical evolution of outlier detection methods
  - Statistical tests
    - Limited (univariate, no mixture model, outliers are rare, only one kind of distribution)
    - No emphasis on computational time
  - Extensions to these tests
    - Multivariate, mixture models, …
    - Still no emphasis on computational time
  - Database-driven approaches
    - First, still statistically driven intuition of outliers
    - Emphasis on computational complexity
  - Database and data mining approaches
    - Spatial intuition of outliers
    - Even stronger focus on computational complexity (e.g. invention of top-*n* problem to propose new efficient algorithms)

Summary

- Consequence
  - Different models are based on different assumptions to model outliers
    - These assumptions are often not explicit but only implicit and not well understood
  - Different models provide different types of output (labeling/scoring)
  - Different models consider outlier at different resolutions (global/local)
  - Thus, different models will produce different results
  - A thorough and comprehensive comparison between different models and approaches is still missing

Summary

Outlook

- Experimental evaluation of different approaches to understand and compare differences and common properties
- A first step towards unification of the diverse approaches: providing density-based outlier scores as probability values [Kriegel et al. 2009a]: judging the deviation of the outlier score from the expected value
- Visualization
- New models
- Performance issues
- Complex data types
- High-dimensional data
- …
- **Und v.a. jede Menge offene Themen für DA, MA, BA Arbeiten**

Literature

Achtert, E., Kriegel, H.-P., Reichert, L., Schubert, E., Wojdanowski, R., Zimek, A. 2010. Visual Evaluation of Outlier Detection Models. In Proc. International Conference on Database Systems for Advanced Applications (DASFAA), Tsukuba, Japan.

Aggarwal, C.C. and Yu, P.S. 2000. Outlier detection for high dimensional data. In Proc. ACM SIGMOD Int. Conf. on Management of Data (SIGMOD), Dallas, TX.

Angiulli, F. and Pizzuti, C. 2002. Fast outlier detection in high dimensional spaces. In Proc. European Conf. on Principles of Knowledge Discovery and Data Mining, Helsinki, Finland.

Arning, A., Agrawal, R., and Raghavan, P. 1996. A linear method for deviation detection in large databases. In Proc. Int. Conf. on Knowledge Discovery and Data Mining (KDD), Portland, OR.

Barnett, V. 1978. The study of outliers: purpose and model. Applied Statistics, 27(3), 242–250.

Bay, S.D. and Schwabacher, M. 2003. Mining distance-based outliers in near linear time with randomization and a simple pruning rule. In Proc. Int. Conf. on Knowledge Discovery and Data Mining (KDD), Washington, DC.

Breunig, M.M., Kriegel, H.-P., Ng, R.T., and Sander, J. 1999. OPTICS-OF: identifying local outliers. In Proc. European Conf. on Principles of Data Mining and Knowledge Discovery (PKDD), Prague, Czech Republic.

Breunig, M.M., Kriegel, H.-P., Ng, R.T., and Sander, J. 2000. LOF: identifying density-based local outliers. In Proc. ACM SIGMOD Int. Conf. on Management of Data (SIGMOD), Dallas, TX.

Literature

Ester, M., Kriegel, H.-P., Sander, J., and Xu, X. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In Proc. Int. Conf. on Knowledge Discovery and Data Mining (KDD), Portland, OR.

Fan, H., Zaïane, O., Foss, A., and Wu, J. 2006. A nonparametric outlier detection for efficiently discovering top-n outliers from engineering data. In Proc. Pacific-Asia Conf. on Knowledge Discovery and Data Mining (PAKDD), Singapore.

Ghoting, A., Parthasarathy, S., and Otey, M. 2006. Fast mining of distance-based outliers in high dimensional spaces. In Proc. SIAM Int. Conf. on Data Mining (SDM), Bethesda, ML.

Hautamaki, V., Karkkainen, I., and Franti, P. 2004. Outlier detection using k-nearest neighbour graph. In Proc. IEEE Int. Conf. on Pattern Recognition (ICPR), Cambridge, UK.

Hawkins, D. 1980. Identification of Outliers. Chapman and Hall.

Jin, W., Tung, A., and Han, J. 2001. Mining top-*n* local outliers in large databases. In Proc. ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (SIGKDD), San Francisco, CA.

Jin, W., Tung, A., Han, J., and Wang, W. 2006. Ranking outliers using symmetric neighborhood relationship. In Proc. Pacific-Asia Conf. on Knowledge Discovery and Data Mining (PAKDD), Singapore.

Johnson, T., Kwok, I., and Ng, R.T. 1998. Fast computation of 2-dimensional depth contours. In Proc. Int. Conf. on Knowledge Discovery and Data Mining (KDD), New York, NY.

Knorr, E.M. and Ng, R.T. 1997. A unified approach for mining outliers. In Proc. Conf. of the Centre for Advanced Studies on Collaborative Research (CASCON), Toronto, Canada.

Literature

Knorr, E.M. and NG, R.T. 1998. Algorithms for mining distance-based outliers in large datasets. In Proc. Int. Conf. on Very Large Data Bases (VLDB), New York, NY.

Knorr, E.M. and Ng, R.T. 1999. Finding intensional knowledge of distance-based outliers. In Proc. Int. Conf. on Very Large Data Bases (VLDB), Edinburgh, Scotland.

Kriegel, H.-P., Kröger, P., Schubert, E., and Zimek, A. 2009. Outlier detection in axis-parallel subspaces of high dimensional data. In Proc. Pacific-Asia Conf. on Knowledge Discovery and Data Mining (PAKDD), Bangkok, Thailand.

Kriegel, H.-P., Kröger, P., Schubert, E., and Zimek, A. 2009a. LoOP: Local Outlier Probabilities. In Proc. ACM Conference on Information and Knowledge Management (CIKM), Hong Kong, China.

Kriegel, H.-P., Schubert, M., and Zimek, A. 2008. Angle-based outlier detection, In Proc. ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (SIGKDD), Las Vegas, NV.

McCallum, A., Nigam, K., and Ungar, L.H. 2000. Efficient clustering of high-dimensional data sets with application to reference matching. In Proc. ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (SIGKDD), Boston, MA.

Papadimitriou, S., Kitagawa, H., Gibbons, P., and Faloutsos, C. 2003. LOCI: Fast outlier detection using the local correlation integral. In Proc. IEEE Int. Conf. on Data Engineering (ICDE), Hong Kong, China.

Pei, Y., Zaiane, O., and Gao, Y. 2006. An efficient reference-based approach to outlier detection in large datasets. In Proc. 6th Int. Conf. on Data Mining (ICDM), Hong Kong, China.

Preparata, F. and Shamos, M. 1988. Computational Geometry: an Introduction. Springer Verlag.

Literature

Ramaswamy, S. Rastogi, R. and Shim, K. 2000. Efficient algorithms for mining outliers from large data sets. In Proc. ACM SIGMOD Int. Conf. on Management of Data (SIGMOD), Dallas, TX.

Rousseeuw, P.J. and Leroy, A.M. 1987. Robust Regression and Outlier Detection. John Wiley.

Ruts, I. and Rousseeuw, P.J. 1996. Computing depth contours of bivariate point clouds. Computational Statistics and Data Analysis, 23, 153–168.

Tao Y., Xiao, X. and Zhou, S. 2006. Mining distance-based outliers from large databases in any metric space. In Proc. ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (SIGKDD), New York, NY.

Tan, P.-N., Steinbach, M., and Kumar, V. 2006. Introduction to Data Mining. Addison Wesley.

Tang, J., Chen, Z., Fu, A.W.-C., and Cheung, D.W. 2002. Enhancing effectiveness of outlier detections for low density patterns. In Proc. Pacific-Asia Conf. on Knowledge Discovery and Data Mining (PAKDD), Taipei, Taiwan.

Tukey, J. 1977. Exploratory Data Analysis. Addison-Wesley.

Zhang, T., Ramakrishnan, R., Livny, M. 1996. BIRCH: an efficient data clustering method for very large databases. In Proc. ACM SIGMOD Int. Conf. on Management of Data (SIGMOD), Montreal, Canada.