FastLOF: An Expectation-Maximization based Local Outlier Detection Algorithm

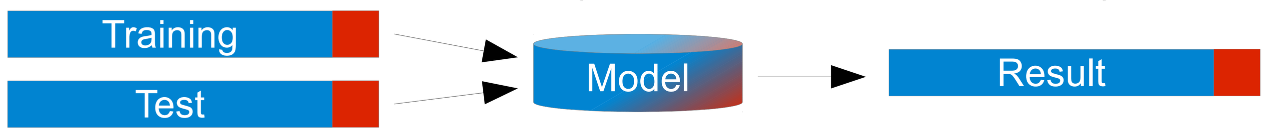
Markus Goldstein

German Research Center for Artificial Intelligence (DFKI), Kaiserslautern

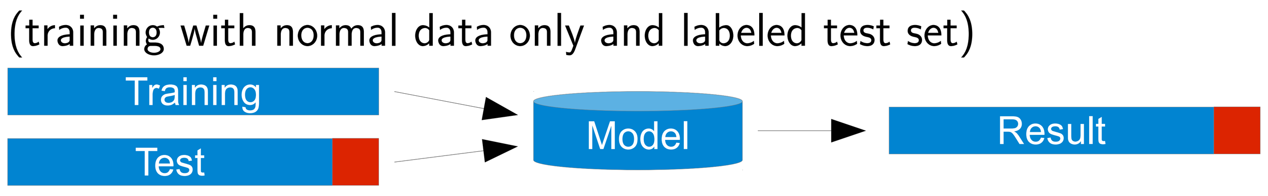
www.dfki.de   markus.goldstein@dfki.de

Introduction

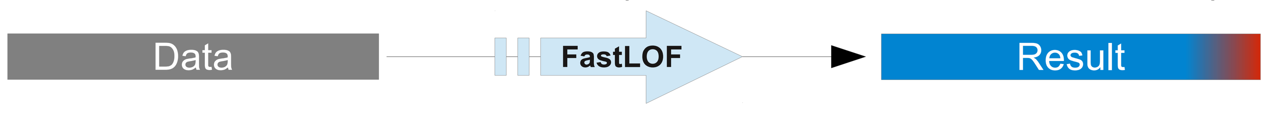
- Anomaly detection finds **outliers** in data sets which
  - only occur very rarely in the data and
  - their features significantly deviate from the normal data
- Three different anomaly detection setups exist [4]:
  1. Supervised anomaly detection (labeled training and test set)



  2. Semi-supervised anomaly detection (training with normal data only and labeled test set)



  3. Unsupervised anomaly detection (one data set without any labels)



- In this work, we present an **unsupervised** algorithm which **scores** instances in a given data set according to their **outlierliness**

Related Work

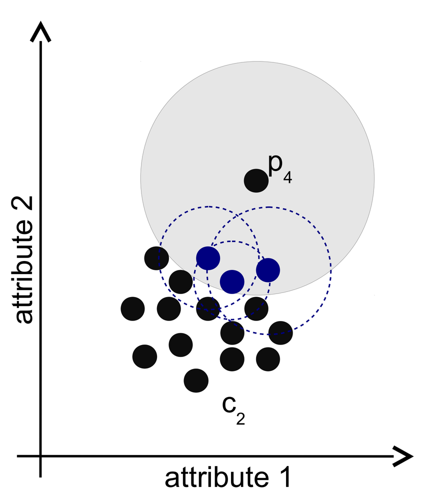
Unsupervised anomaly detection [4]

- Nearest-neighbor based algorithms
  - Global k-nearest-neighbor (k-NN) [9]
  - Well known local method: Local Outlier Factor (LOF) [3]
  - Many improvements based on LOF: Connectivity-Based Outlier Factor (COF) [10], Local Outlier Probability (LoOP) [7], Influenced Outlierness (INFLO) [6] and Local Correlation Integral (LOCI) [8]
  - Best performing methods today [2]
  - Computational effort for nearest-neighbor search basically $O(n^2)$
- Clustering based algorithms
  - Use k-means to cluster the data first
  - Compute CBLOF [5] or LDCOF [1] scores based on clustering results
  - Can be faster than k-NN methods
- Statistical methods
  - Parametric methods, e.g. Gaussian Mixture Models (GMM)
  - Non-parametric methods, e.g. histograms or kernel-density estimation (KDE)

Local Outlier Factor (LOF)

- Introduced by Breunig et al in 2000 [3]
- Three steps to compute LOF score:
  1. Find the k-nearest-neighbors
  2. For each instance compute the local reachability density:

$$LRD_{min}(p) = 1 / \left(\frac{\sum_{o \in N_{min}(p)} reach\_dist_{min}(p, o)}{|N_{min}(p)|} \right)$$



  3. For each instance compute the ratio of local densities

$$LOF_{min}(p) = \frac{\sum_{o \in N_{min}(p)} \frac{LRD_{min}(o)}{LRD_{min}(p)}}{|N_{min}(p)|}$$

- Scores close to 1.0 indicate normal data
- Scores $>$ (1.2 ... 2.0) are anomalies
- Most computational effort is finding the nearest neighbors $O(n^2)$, often $>$ 99% of the run time

Acknowledgment: This work is part of ADEWaS, a project of Deutsche Telekom Laboratories.

Performance Improvement Attempts

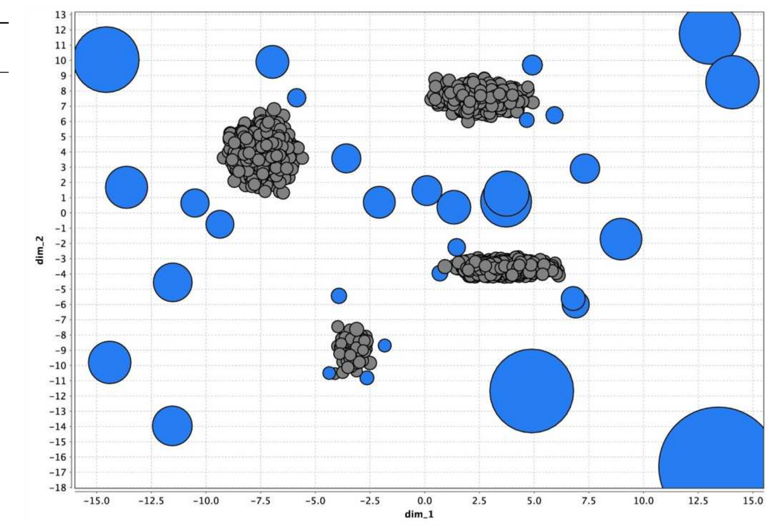
- Space partitioning algorithms (e.g. search trees): Require time to build the tree structure and can be slow when having many dimensions
- Locality Sensitive Hashing (LSH): Approximates neighbors well in dense areas but performs poor for outliers

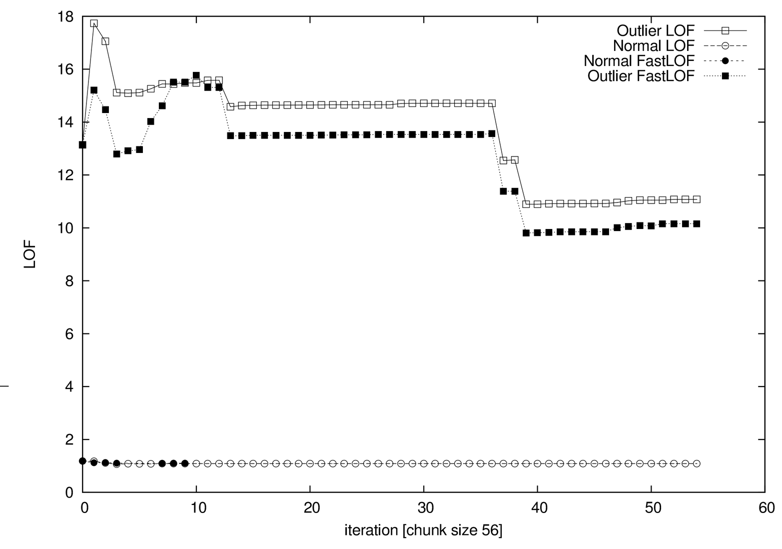
FastLOF

- Idea: Estimate the nearest neighbors for dense areas approximately and compute exact neighbors for sparse areas
- Expectation step: Find some (approximately correct) neighbors and estimate LRD/LOF based on them
- Maximization step: For promising candidates (LOF $> \theta$), find better neighbors

Algorithm 1 The FastLOF algorithm

```
Input
    D = d_1, ..., d_n: data set with N instances
    c: chunk size (e.g. √N)
    θ: threshold for LOF
    k: number of nearest neighbors
Output
    LOF = lof_1, ..., lof_n: estimated LOF scores
function FastLOF(D, c, θ, k)
    shuffle(D)
    Group d_1, ..., d_n in chunk_1, ..., chunk_c
    active ← D
    while new NN^k found do
        for all d_i ∈ active do
            NN_i^k ← findNN(d_i, chunk_{c_i})
            Update NN_x^k for new neighbor x in NN_i^k
            c_i++
        LRD ← LRD(D, NN^k)
        LOF ← LOF(D, NN^k)
        active ← 0
        for all d_i ∈ D do
            if lof_i > θ then
                active ← d_i
    return LOF
```





Evaluation and Results

- Evaluation using UCI machine learning data sets (preprocessed as in [1]):
  - Breast Cancer Wisconsin data set
  - Pen-Based Recognition of Handwritten Digits data set (global and local anomaly detection task)

Dataset	k	θ	LOF AUC	FastLOF AUC	FastLOF Calcs	Best Alg.	Best AUC	Worst AUC
Breast Cancer Wisconsin	10	1.10	0.9916	0.9882	18,5%	INFLO	0.9922	0.8389
Pen-based 4-anomaly (local)	10	1.01	0.9878	0.9937	16.0%	**FastLOF**	0.9937	0.7010
Pen-based 8-normal (global)	40	1.00	0.8864	0.9050	35.5%	uCBLOF	0.9923	0.6808

- 65% - 80% less distance computations than LOF
- Scores already available as approximations during calculation
- FastLOF scores converge to LOF scores (if θ decreases over time)

References

[1] Mennatallah Amer. Comparison of unsupervised anomaly detection techniques. Bachelor's Thesis, 2011. http://madm.dfki.de/_media/theses/bachelorthesis-amer_2011.pdf.

[2] Mennatallah Amer and Markus Goldstein. Nearest-neighbor and clustering based anomaly detection algorithms for rapidminer. In Ingo Mierswa Simon Fischer, editor, *Proceedings of the 3rd RapidMiner Community Meeting and Conferernce (RCOMM 2012)*, pages 1–12. Shaker Verlag GmbH, 8 2012.

[3] Markus M. Breunig, Hans-Peter Kriegel, Raymond T. Ng, and Jörg Sander. Lof: identifying density-based local outliers. *SIGMOD Rec.*, 29(2):93–104, 2000.

[4] Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey. *ACM Comput. Surv.*, 41(3):1–58, 2009.

[5] Zengyou He, Xiaofei Xu, and Shengchun Deng. Discovering cluster-based local outliers. *Pattern Recognition Letters*, 24(9-10):1641 – 1650, 2003.

[6] Wen Jin, Anthony Tung, Jiawei Han, and Wei Wang. Ranking outliers using symmetric neighborhood relationship. In Wee-Keong Ng, Masaru Kitsuregawa, Jianzhong Li, and Kuiyu Chang, editors, *Advances in Knowledge Discovery and Data Mining*, volume 3918 of *Lecture Notes in Computer Science*, pages 577–593. Springer Berlin / Heidelberg, 2006.

[7] Hans-Peter Kriegel, Peer Kröger, Erich Schubert, and Arthur Zimek. Loop: local outlier probabilities. In *CIKM '09: Proceeding of the 18th ACM conference on Information and knowledge management*, pages 1649–1652, New York, NY, USA, 2009. ACM.

[8] Spiros Papadimitriou, Hiroyuki Kitagawa, Phillip B. Gibbons, and Christos Faloutsos. Loci: Fast outlier detection using the local correlation integral. *Data Engineering, International Conference on*, 0:315, 2003.

[9] Sridhar Ramaswamy, Rajeev Rastogi, and Kyuseok Shim. Efficient algorithms for mining outliers from large data sets. In *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, SIGMOD '00, pages "427–438", New York, NY, USA, 2000. ACM.

[10] Jian Tang, Zhixiang Chen, Ada Fu, and David Cheung. Enhancing effectiveness of outlier detections for low density patterns. In Ming-Syan Chen, Philip Yu, and Bing Liu, editors, *Advances in Knowledge Discovery and Data Mining*, volume 2336 of *Lecture Notes in Computer Science*, pages 535–548. Springer Berlin / Heidelberg, 2002.