Towards Efficient Outlier Detection

Ye Wang, Srinivasan Parthasarathy (Advisor)
[ wangy, srini ]@cse.ohio-state.edu
Department of computer Science and Engineering,
The Ohio State University

Motivation
- Distance based Outlier detection has gained attention from internet security applications to bank fraud analysis.
- However, there are still many aspects to desire in the algorithm efficiency.
- There are various optimizations to improve the performance. Which optimizations are better?
- What are the current bottlenecks of the detection algorithms?
- What can we learn from the design of the existing algorithms? Can we infer some guidelines for designing better outlier detection algorithms?

Distance Based algorithms: State of Art
- ORCA [BS01], RB-PRG [GPO06], Nested Loop algorithm [KN99]
- Approximate Nearest Neighbor Searching (ANNS) [RRS00, GPO06]
- Use partitioning algorithms for preprocessing [RRS00, GPO06, VG09] – K-Means, Bisection algorithm, BIRCH
- Pruning partitions when searching neighbors [RRS00, VG09]

Various Optimizations – A glance
- Approximate Nearest Neighbor Search (ANNS)
- Partitioning for Preprocessing
- Partition Pruning while Searching for Neighbors (PPSN)
- Rank Object Candidates while searching for Neighbors (ROCN)
- Rank Object Candidates while searching for Outliers (ROCO)
- ROCO with partitions

Partitioning for Preprocessing
- Partition database and generate partition statistics for each partition. The partition statistics can be used for various kinds of pruning and ranking.

Partition Pruning while Searching for Neighbors (PPSN)
- When searching for top K near neighbors, prune partitions that are far away from the object by utilizing the partition statistics.

Rank Object Candidates while searching for Neighbors (ROCN)
- ROCN: While searching the neighbors of object, search close partitions first, as they are more likely to contain neighbors.

Approximate Nearest Neighbor Search (ANNS)
- If the K-th nearest neighbor of point p is smaller than the current weakest outlier, then p is not an outlier.

Current Weakest Outlier Score = 5, so point p is not an outlier.

In the above case, when k = 4, point p can be pruned without seeing all the points.

Categorization on Current Algorithms
- Integrated framework focusing on partitions for sake of pruning and ranking points, also applying the ANNS rule
- It includes all the detecting components: Partitioning, PPSO, PPSN, ROCN and ROCO, ROCO with partitions
- We will use this framework to evaluate the effectiveness of each optimization

Conclusion
- Exploited the design space of detection algorithms
- New strategy for ranking outliers, and process the potential outliers first
- Unified Framework for Distance based outlier detection algorithm
- Factorial analysis on different optimizations
- This is joint work with Gustavo H. Orair, Carlos Teixeira and Wagner Meira Jr., Department of Computer Science, UFMG, Brazil

Future Directions and References
- Still room for better outlier ranking techniques, such as ranking using locality sensitive based scheme
- More suitable clustering algorithm for outlier detection
- Theoretical framework for different outlier detection algorithms
- References

Part 2 – Factorize each component

Consolidation and Renewed Bearing
- ANNS is very effective and should be used as a basic component for the outlier detection
- PPSN is important, as comparing with neighboring point takes major portion of the execution time
- PPSO does not improve the performance much
- No single best optimizations
- Some better ranking scheme, such as the locality sensitive based ranking, could further improve the performance.

DIODE – Framework for Distance-based Outlier DEtection
- Integrated framework focusing on partitions for sake of pruning and ranking points, also applying the ANNS rule
- It includes all the detecting components: Partitioning, PPSO, PPSN, ROCN and ROCO, ROCO with partitions
- We will use this framework to evaluate the effectiveness of each optimization

Future Directions and References
- Still room for better outlier ranking techniques, such as ranking using locality sensitive based scheme
- More suitable clustering algorithm for outlier detection
- Theoretical framework for different outlier detection algorithms
- References

Part 2 – Factorize each component

Consolidation and Renewed Bearing
- ANNS is very effective and should be used as a basic component for the outlier detection
- PPSN is important, as comparing with neighboring point takes major portion of the execution time
- PPSO does not improve the performance much
- No single best optimizations
- Some better ranking scheme, such as the locality sensitive based ranking, could further improve the performance.

DIODE – Framework for Distance-based Outlier DEtection
- Integrated framework focusing on partitions for sake of pruning and ranking points, also applying the ANNS rule
- It includes all the detecting components: Partitioning, PPSO, PPSN, ROCN and ROCO, ROCO with partitions
- We will use this framework to evaluate the effectiveness of each optimization

Future Directions and References
- Still room for better outlier ranking techniques, such as ranking using locality sensitive based scheme
- More suitable clustering algorithm for outlier detection
- Theoretical framework for different outlier detection algorithms
- References

Part 2 – Factorize each component

Consolidation and Renewed Bearing
- ANNS is very effective and should be used as a basic component for the outlier detection
- PPSN is important, as comparing with neighboring point takes major portion of the execution time
- PPSO does not improve the performance much
- No single best optimizations
- Some better ranking scheme, such as the locality sensitive based ranking, could further improve the performance.

DIODE – Framework for Distance-based Outlier DEtection
- Integrated framework focusing on partitions for sake of pruning and ranking points, also applying the ANNS rule
- It includes all the detecting components: Partitioning, PPSO, PPSN, ROCN and ROCO, ROCO with partitions
- We will use this framework to evaluate the effectiveness of each optimization

Future Directions and References
- Still room for better outlier ranking techniques, such as ranking using locality sensitive based scheme
- More suitable clustering algorithm for outlier detection
- Theoretical framework for different outlier detection algorithms
- References