clustering based algorithm, LDCOF with k-means was used. The algorithm was started once with 30 random centroids. Using 10 optimization steps, an average run time of 20.0ms was achieved, with 100 optimization steps, which was our default setting for the performance comparison, the algorithm took 30.0ms. We expect clustering based methods to be much faster than nearest-neighbor based algorithms on larger data sets. However, HBOS was significantly faster than both: It took 3.8ms with dynamic bin widths and 4.1ms using a fixed bin width. Thus, in our experiments HBOS was 7 times faster than nearest-neighbor based methods and 5 times faster than the k-means based LDCOF. On larger data sets the speed-up can be much higher: On a not publicly available data set comprising of 1,000,000 instances with 15 dimensions, LOF took 23 hours and 46 minutes whereas HBOS took 38 seconds only (dynamic bin-width: 46 seconds).

5 Conclusion

In this paper we present an unsupervised histogram-based outlier detection algorithm (HBOS), which models univariate feature densities using histograms with a fixed or a dynamic bin width. Afterwards, all histograms are used to compute an anomaly score for each data instance. Compared to other algorithms, HBOS works in linear time $O(n)$ in case of fixed bin width or in $O(n \cdot log(n))$ using dynamic bin widths. The evaluation shows that HBOS performs well on global anomaly detection problems but cannot detect local outliers. A comparison of run times also show that HBOS is much faster than standard algorithms, especially on large data sets.

References

1. Amer, M.: Comparison of unsupervised anomaly detection techniques. Bachelor's Thesis 2011, `http://www.madm.eu/_media/theses/thesis-amer.pdf`
2. Amer, M., Goldstein, M.: Nearest-neighbor and clustering based anomaly detection algorithms for rapidminer. In: Proc. of the 3rd RCOMM 2012
3. Breunig, M.M., Kriegel, H.P., Ng, R.T., Sander, J.: Lof: identifying density-based local outliers. SIGMOD Rec. 29(2), 93–104 (2000)
4. Chandola, V., Banerjee, A., Kumar, V.: Anomaly detection: A survey. ACM Comput. Surv. 41(3), 1–58 (2009)
5. Goldstein, M.: FastLOF: An expectation-maximization based local outlier detection algorithm. In: Proc. of the Int. Conf. on Pattern Recognition (2012)
6. He, Z., Xu, X., Deng, S.: Discovering cluster-based local outliers. Pattern Recognition Letters 24(9-10), 1641 – 1650 (2003)
7. Kim, Y., Lau, W.C., et al: Packetscore: statistics-based overload control against distributed denial-of-service attacks. In: INFOCOM 2004. vol. 4, pp. 2594 – 2604
8. Kind, A., Stoecklin, M., Dimitropoulos, X.: Histogram-based traffic anomaly detection. Network and Service Management, IEEE Transactions on 6(2), 110 –121
9. Mierswa, I., Wurst, M., et al: Yale (now: Rapidminer): Rapid prototyping for complex data mining tasks. In: Proc. of the ACM SIGKDD 2006
10. Papadimitriou, S., Kitagawa, H., et al: Loci: Fast outlier detection using the local correlation integral. Int. Conf. on Data Engineering p. 315 (2003)
11. Ramaswamy, S., Rastogi, R., Shim, K.: Efficient algorithms for mining outliers from large data sets. pp. "427–438". SIGMOD '00