Chapter 12

Outlier Detection

12.1 Bibliographic Notes

Hawkins [Haw80] defined outliers from a statistics angle. For surveys or tutorials on the subject of outlier and anomaly detection, see [CBK09, HA04, ABA06, MS03a, MS03b, PP07, BC83, BG05, BMAD06]. Song, Wu, Jermaine, et al. [SWJR07] proposed the notion of conditional anomaly and contextual outlier detection.

Fujimaki, Yairi, and Machida [FYM05] presented an example of semi-supervised outlier detection using a set of labeled "normal objects". For an example of semi-supervised outlier detection using labeled outliers, see [DM02].

Shewhart [She31] assumed that most objects follow a Gaussian distribution and used 3σ as the threshold for identifying outliers, where σ is the standard deviation. Boxplots are used to detect and visualize outliers in various applications such as medical data [HFLP01]. Grubb's test was described by Grubbs [Gru69], Stefansky [Ste72], and Anscombe and Guttman [AG60]. Laurikkala, Juhola, and Kentala [LJK00] and Aggarwal and Yu [AY01] extended the Grubb's test to detect multivariate outliers. Use of the χ^2-statistic to detect multivariate outliers was conducted by Ye and Chen [YC01].

Agarwal [Aga06] used Gaussian mixture models to capture "normal data". Abraham and Box [AB79] assumed that outliers are generated by a normal distribution with a substantially larger variance. Eskin [Esk00] used the EM algorithm to learn mixture models for "normal data" and outliers.

Histogram-based outlier detection methods are popular in the application domain of intrusion detection [Esk00, EAP$^+$02] and fault detection [FP97].

The notion of distance-based outliers was developed by Knorr and Ng [KN97]. The index-based, nested-loop based, and grid-based approaches were explored [KN98, KNT00] to speed up distance-based outlier detection. Bay and Schwabacher [BS03] pointed out that the CPU runtime of the nested-loop method is often scalable with respect to the database size. Tao, Xiao, and Zhou [TXZ06] presented an algorithm that finds all distance-based outliers by scanning the database three times with fixed main memory. When the memory size is larger, they proposed a method that uses only one or two scans.

The notion of density-based outliers was firstly developed by Breunig, Kriegel, Ng, and Sander [BKNS00]. Various methods proposed under the theme of density-based outlier detection include [JTH01, JTHW06, PKGF03]. The variations differ in how they estimate density.

The bootstrap method discussed in Example 12.17 was developed by Barbara, Li, and Couto et al. [BLC+03]. The FindCBOLF algorithm was given by He, Xu, and Deng [HXD03]. For the use of fixed-width clustering in outlier detection methods, see [EAP+02, MC03, HXD03]. Barbara, Wu, and Jajodia [BWJ01] used multi-class classification in network intrusion detection.

Song, Wu, Jermaine, et al. [SWJR07] and Fawcet and Provost [FP97] presented a method to reduce the problem of contextual outlier detection to conventional outlier detection. Yi, Sidiropoulos, Johnson, Jagadish et al. [YSJ+00] used regression techniques to detect contextual outliers in co-evolving sequences. The idea in Example 12.22 for collective outlier detection on graph data is based on Noble and Cook [NC03].

The HilOut algorithm was proposed by Angiulli and Pizzuti [AP05]. Aggarwal and Yu [AY01] developed the sparsity coefficient-based subspace outlier detection method. Kriegel, Schubert, and Zimek [KSZ08] proposed angle-based outlier detection.

Bibliography

[AB79] B. Abraham and G.E.P. Box. Bayesian analysis of some outlier problems in time series. *Biometrika*, 66:229–248, 1979.

[ABA06] M. Agyemang, K. Barker, and R. Alhajj. A comprehensive survey of numeric and symbolic outlier mining techniques. *Intell. Data Anal.*, 10:521–538, 2006.

[AG60] F. J. Anscombe and I. Guttman. Rejection of outliers. *Technometrics*, 2:123–147, 1960.

[Aga06] D. Agarwal. Detecting anomalies in cross-classified streams: a bayesian approach. *Knowl. Inf. Syst.*, 11:29–44, 2006.

[AP05] F. Angiulli and C. Pizzuti. Outlier mining in large high-dimensional data sets. *IEEE Trans. on Knowl. and Data Eng.*, 17:203–215, 2005.

[AY01] C. C. Aggarwal and P. S. Yu. Outlier detection for high dimensional data. In *Proc. 2001 ACM-SIGMOD Int. Conf. Management of Data (SIGMOD'01)*, pages 37–46, Santa Barbara, CA, May 2001.

[BC83] R.J. Beckman and R.D. Cook. Outlier...s. *Technometrics*, 25:119–149, 1983.

[BG05] I. Ben-Gal. Outlier detection. In *O. Maimon and L. Rockach (eds.) Data Mining and Knowledge Discovery Handbook: A Complete Guide for Practitioners and Researchers*, Kluwer Academic, 2005.

[BKNS00] M. M. Breunig, H.-P. Kriegel, R. Ng, and J. Sander. LOF: Identifying density-based local outliers. In *Proc. 2000 ACM-SIGMOD Int. Conf. Management of Data (SIGMOD'00)*, pages 93–104, Dallas, TX, May 2000.

[BLC+03] D. Barbará, Y. Li, J. Couto, J.-L. Lin, and S. Jajodia. Bootstrapping a data mining intrusion detection system. In *Proc. 2003 ACM Symp. Applied Computing (SAC'03)*, March 2003.

[BMAD06] Z. A. Bakar, R. Mohemad, A. Ahmad, and M. M. Deris. A comparative study for outlier detection techniques in data mining. In *Proc. 2006 IEEE Conf. Cybernetics and Intelligent Systems*, pages 1–6, Bangkok, Thailand, 2006.

[BS03] S. D. Bay and M. Schwabacher. Mining distance-based outliers in near linear time with randomization and a simple pruning rule. In *Proc. 2003 ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining (KDD'03)*, pages 29–38, Washington, DC, Aug 2003.

[BWJ01] D. Barbara, N. Wu, and S. Jajodia. Detecting novel network intrusion using bayesian estimators. In *Proc. 2001 SIAM Int. Conf. Data Mining (SDM'01)*, Chicago, IL, April 2001.

[CBK09] V. Chandola, A. Banerjee, and V. Kumar. Anomaly detection: A survey. *ACM Computing Surveys*, 41:1–58, 2009.

[DM02] D. Dasgupta and N.S. Majumdar. Anomaly detection in multidimensional data using negative selection algorithm. In *Proc. 2002 Congress on Evolutionary Computation (CEC'02)*, pages 1039–1044, Washington DC, 2002.

[EAP+02] E. Eskin, A. Arnold, M. Prerau, L. Portnoy, and S. Stolfo. A geometric framework for unsupervised anomaly detection: Detecting intrusions in unlabeled data. In *Proc. 2002 Int. Conf. of Data Mining for Security Applications*, 2002.

[Esk00] E. Eskin. Anomaly detection over noisy data using learned probability distributions. In *Proc. 17th Int. Conf. Machine Learning (ICML'00)*, 2000.

[FP97] T. Fawcett and F. Provost. Adaptive fraud detection. *Data Mining and Knowledge Discovery*, 1:291–316, 1997.

[FYM05] R. Fujimaki, T. Yairi, and K. Machida. An approach to spacecraft anomaly detection problem using kernel feature space. In *Proc. 2005 Int. Workshop on Link Discovery (LinkKDD'05)*, pages 401–410, Chicago, Illinois, 2005.

[Gru69] F. E. Grubbs. Procedures for detecting outlying observations in samples. *Technometrics*, 11:1–21, 1969.

[HA04] V. J. Hodge and J. Austin. A survey of outlier detection methodologies. *Artificial Intelligence Review*, 22:85–126, 2004.

[Haw80] D. M. Hawkins. *Identification of Outliers*. Chapman and Hall, London, 1980.

[HFLP01] P. S. Horn, L. Feng, Y. Li, and A. J. Pesce. Effect of outliers and nonhealthy individuals on reference interval estimation. *Clinical Chemistry*, 47:2137–2145, 2001.

[HXD03] Z. He, X. Xu, and S. Deng. Discovering cluster-based local outliers. *Pattern Recognition Lett.*, 24:1641–1650, June, 2003.

[JTH01] W. Jin, A.K.H. Tung, and J. Han. Mining top-n local outliers in large databases. In *Proc. 2001 ACM SIGKDD Int. Conf. Knowledge Discovery in Databases (KDD'01)*, pages 293–298, San Fransisco, CA, Aug. 2001.

[JTHW06] W. Jin, A.K.H. Tung, J. Han, and W. Wang. Ranking outliers using symmetric neighborhood relationship. In *Proc. 2006 Pacific-Asia Conf. Knowledge Discovery and Data Mining (PAKDD'06)*, Singapore, April 2006.

[KN97] E. Knorr and R. Ng. A unified notion of outliers: Properties and computation. In *Proc. 1997 Int. Conf. Knowledge Discovery and Data Mining (KDD'97)*, pages 219–222, Newport Beach, CA, Aug. 1997.

[KN98] E. Knorr and R. Ng. Algorithms for mining distance-based outliers in large datasets. In *Proc. 1998 Int. Conf. Very Large Data Bases (VLDB'98)*, pages 392–403, New York, NY, Aug. 1998.

[KNT00] E. M. Knorr, R. T. Ng, and V. Tucakov. Distance-based outliers: Algorithms and applications. *The VLDB J.*, 8:237–253, 2000.

[KSZ08] H.-P. Kriegel, M. Schubert, and A. Zimek. Angle-based outlier detection in high-dimensional data. In *Proc. 2008 ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD'08)*, pages 444–452, Las Vegas, NV, Aug. 2008.

[LJK00] J. Laurikkala, M. Juhola, and E. Kentala. Informal identification of outliers in medical data. In *Proc. 5th Int. Workshop on Intelligent Data Analysis in Medicine and Pharmacology*, pages 20–24, Berlin, Germany, Aug. 2000.

[MC03] M. V. Mahoney and P. K. Chan. Learning rules for anomaly detection of hostile network traffic. In *Proc. 2003 Int. Conf. Data Mining (ICDM'03)*, Melbourne, FL, Nov. 2003.

[MS03a] M. Markou and S. Singh. Novelty detection: A review—part 1: Statistical approaches. *Signal Processing*, 83:2481–2497, 2003.

[MS03b] M. Markou and S. Singh. Novelty detection: A review—part 2: Neural network based approaches. *Signal Processing*, 83:2499–2521, 2003.

[NC03] C. C. Noble and D. J. Cook. Graph-based anomaly detection. In *Proc. 2003 ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining (KDD'03)*, pages 631–636, Washington, DC, Aug 2003.

[PKGF03] S. Papadimitriou, H. Kitagawa, P. B. Gibbons, and C. Faloutsos. Loci: Fast outlier detection using the local correlation integral. In *Proc. 2003 Int. Conf. Data Engineering (ICDE'03)*, pages 315–326, Bangalore, India, March 2003.

[PP07] A. Patcha and J.-M. Park. An overview of anomaly detection techniques: Existing solutions and latest technological trends. *Comput. Netw.*, 51, 2007.

[She31] W. A. Shewhart. *Economic Control of Quality of Manufactured Product.* D. Van Nostrand Company, 1931.

[Ste72] W. Stefansky. Rejecting outliers in factorial designs. *Technometrics*, 14:469–479, 1972.

[SWJR07] X. Song, M. Wu, C. Jermaine, and S. Ranka. Conditional anomaly detection. *IEEE Trans. on Knowl. and Data Eng.*, 19, 2007.

[TXZ06] Y. Tao, X. Xiao, and S. Zhou. Mining distance-based outliers from large databases in any metric space. In *Proc. 2006 ACM SIGKDD Int. Conf. Knowledge Discovery in Databases (KDD'06)*, pages 394–403, Philadelphia, PA, Aug. 2006.

[YC01] N. Ye and Q. Chen. An anomaly detection technique based on a chi-square statistic for detecting intrusions into information systems. *Quality and Reliability Engineering International*, 17:105–112, 2001.

[YSJ+00] B.-K. Yi, N. Sidiropoulos, T. Johnson, H. V. Jagadish, C. Faloutsos, and A. Biliris. Online data mining for co-evolving time sequences. In *Proc. 2000 Int. Conf. Data Engineering (ICDE'00)*, pages 13–22, San Diego, CA, Feb. 2000.