Encyclopedia of Business Analytics and Optimization

John Wang
Montclair State University, USA

Volume IV
Op-So

BUSINESS SCIENCE Reference
An Imprint of IGI Global

Managing Director:	Lindsay Johnston
Production Editor:	Jennifer Yoder
Development Editor:	Austin DeMarco
Acquisitions Editor:	Kayla Wolfe
Typesetter:	Christina Barkanic, Michael Brehm, John Crodian, Lisandro Gonzalez, Christina Henning, Deanna Jo Zombro
Cover Design:	Jason Mull

Published in the United States of America by
Business Science Reference (an imprint of IGI Global)
701 E. Chocolate Avenue
Hershey PA 17033
Tel: 717-533-8845
Fax: 717-533-8661
E-mail: cust@igi-global.com
Web site: http://www.igi-global.com

Library of Congress Cataloging-in-Publication Data

Encyclopedia of business analytics and optimization / John Wang, editor.
pages cm
Includes bibliographical references and index.
Summary: "This reference confronts the challenges of information retrieval in the age of Big Data by exploring recent advances in the areas of knowledge management, data visualization, interdisciplinary communication, and others"-- Provided by publisher.
ISBN 978-1-4666-5202-6 (hardcover) -- ISBN 978-1-4666-5203-3 (ebook) -- ISBN 978-1-4666-5205-7 (print & perpetual access) 1. Management--Mathematical models. 2. Decision making--Mathematical models. 3. Business planning--Mathematical models. 4. Big data. I. Wang, John, 1955-
HD30.25.E53 2014
658.4'038--dc23

2013046204

British Cataloguing in Publication Data
A Cataloguing in Publication record for this book is available from the British Library.

Outlier Detection in Big Data

Victoria J. Hodge
University of York, UK

INTRODUCTION

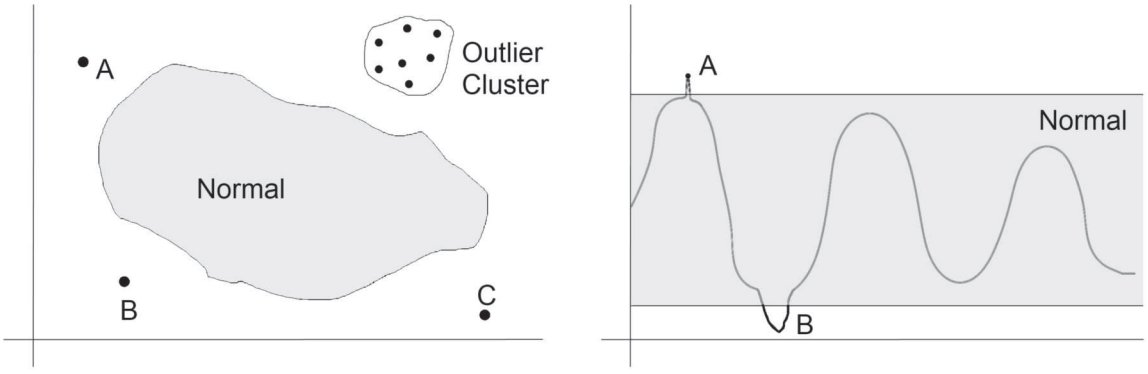
This chapter will examine the issues posed by Big Data for the task of outlier detection. An outlier (Hodge, 2011) (often called an anomaly (Chandola, Banerjee, & Kumar, 2009) in the literature) is a particular data point or, in some instances, a small set of data points that is inconsistent with the rest of the data population as shown in Figure 1.

"Big Data" refers to large, dynamic collections of data. Data sources are generating more and more data while increasing numbers of decentralized data sources are added everyday as interconnection and data exchange become easier. Typical features of Big Data are: data comprising trillions of records where the data is loosely structured; delivered from heterogeneous data sources in heterogeneous data formats; often streamed in real-time and at high volume; and, often distributed either across local computer clusters or across separate geographically distinct sites driven by Big Data mechanisms such as cloud computing and on-line services. Such data may be problematic for traditional outlier tools and techniques to process. This chapter studies when and where outlier detection is used and examines the problems posed and the solutions produced for outlier detection on Big Data. It then analyzes the future directions for outlier detection in Big Data.

BACKGROUND

Outlier detection or anomaly detection has been used for centuries to detect and remove anomalous data points from data. The original methods were arbitrary but today, principled and systematic techniques are used. These include (Hodge, 2011): distance-based; density-based; statistical (including regression); machine learning (including decision trees, expert systems and clustering); information theory; spectral decomposition; neural networks; support vector machines (SVMs); and, natural computation derived from artificial immune systems. Outlier detection distinguishes outlier data from normal data using either: abnor-

Figure 1. The graph on the left includes three outliers (A-C) and a small cluster of outliers. The graph on the right represents time-series data with a single point outlier (A) and an outlying section (B).



DOI: 10.4018/978-1-4666-5202-6.ch157

mality detection which compares new data to a model of normality (or a model of abnormality); or, outlier classification which classifies new data as either normal or abnormal. Outlier detection can also use time-series or sequence analysis to detect changes in temporal patterns.

In the business domain, outlier detection can rapidly identify an intruder inside a business's computer network with malicious intentions (Vieira, Schulter, Westphall, & Westphall, 2010). DARPA (http://www.darpa.mil) is investing $35 million in a program focusing on insider threat detection in massive datasets as anomaly detection produces important information for a wide variety of application domains. Much outlier detection research focuses on detecting fraud, particularly financial fraud (Phua, Lee, Smith-Miles, & Gayler, 2010). Fraud detection automates all or part of the application process and the usage or activity monitoring. In general business databases, outliers may indicate fraudulent cases or they may just denote an error. Outlier detection can pinpoint these data so they can be corrected or removed and database consistency and integrity can be ensured. Equity or commodity traders can use outlier detection to monitor individual shares, commodities or markets to detect buying or selling opportunities (Fang, Luo, Xu, & Fei, 2009). Businesses can identify new opportunities by using outlier detection to pinpoint unusual or distinctive patents using text-based outlier detection (Yoon & Kim, 2012). Outlier detection can even be used to provide an early warning to detect financial institutions that display abnormal behavior and may be more likely to fail (Kimmel, Booth, & Booth, 2010). Activity monitoring of time-series or sequence data can be used to constantly monitor processes for anomalies: detecting faults in machinery (Schlechtingen & Santos, 2011), detecting faults on factory production lines (Merdan, Vallee, Lepuschitz, & Zoitl, 2011) or analyzing telecommunication networks (Eiweck, Pattinson, Behringer, & Seewald, 2010). Such fault detection can help to minimize downtime, prevent failures and save businesses money and time. Businesses rely on the transportation systems to transport their products or to receive raw materials. Employees rely on the transport network to get to and from work and to meetings. Hence, an efficient and reliable transportation system is vital for business productivity. Traffic incidents, vehicle defects or infrastructure defects can be detected by processing the sensor data and recognizing outliers.

FINDING OUTLIERS IN BIG DATA

Issues, Controversies, Problems

As the complexity, variety, speed and volume of data increases then management and processing of these data becomes ever more complex. Additionally, many businesses require real-time outlier detection on such data. Hence, outlier detectors need to be carefully designed to cope with the complexity, variety, speed and volume required. The volume of outliers detected in Big Data may well overwhelm many system administrators and software management tools used for diagnosis and analysis. Hence, outlier detectors need to be accurate and minimize false positives or false negatives due to the cost of analyzing each anomaly. The granularity of Big Data needs to be sufficiently high to allow the individual points to be differentiated for outlier analysis. However, Big Data are often very high dimensional. This high dimensionality causes the data points to become sparse so existing distance measures such as Euclidean distance and the standard concept of nearest neighbors become less applicable (Ertoz, Steinbach, & Kumar, 2003). Additional data dimensions can also introduce noise and make outliers more difficult to detect. Outlier detection, therefore, needs to handle high dimensional and sparse data. If this data is distributed, there is also the issue of data synchronization when aggregating the data.

While Big Data poses many challenges for outlier detection applications, it also provides opportunities. Big Data will contain a broader

range of outlier examples and will allow systems to uncover new types of outlier through increased data richness. Outlier detection systems need to find rare patterns (outliers) and can improve their robustness by exploiting the richness of the data as models are only as good as the data learned. Many interesting data patterns and data relationships are not stored in a single data collection but are spread across heterogeneous data sources (Das, Srivastava, Matthews, & Oza, 2010). Businesses must develop new techniques for analysis and visualization of these distributed data sources which will allow new patterns and relationships to be discovered that were not previously available. This will close the gap between what information the data holds and what we can extract. There is also potential to incorporate new data sources. Social network analysis is being introduced to uncover new outliers (Šubelj, Furlan, & Bajec, 2011) or to provide additional evidence (Phua, Smith-Miles, Lee, & Gayler, 2012) for outlier validation.

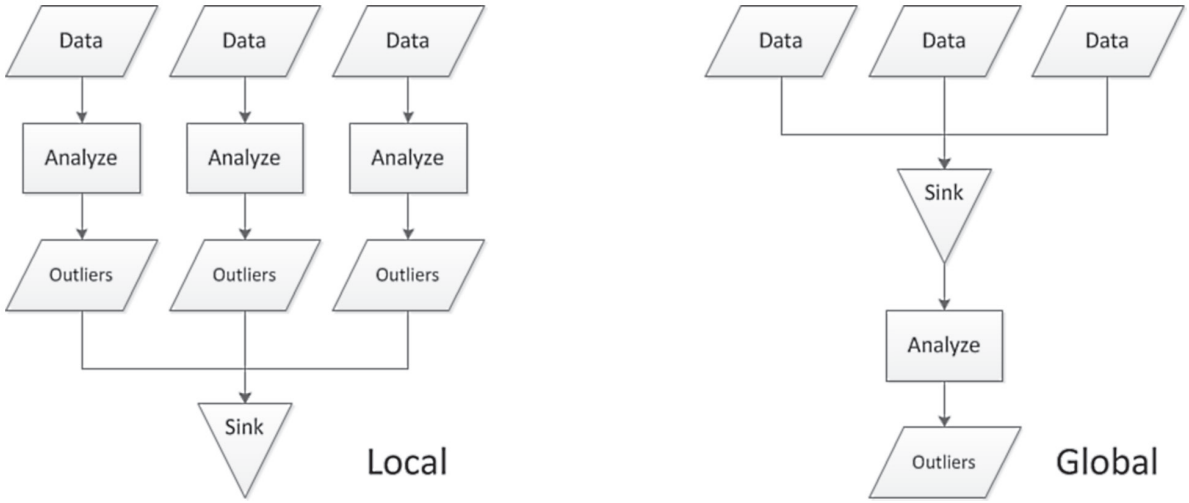
Solutions and Recommendations

Many businesses require real-time decision making including real-time outlier detection. This requires systems that can process vast volumes of data, often heterogeneous streaming data and provide instantaneous decisions. There are various solutions for outlier detection to allow large volumes of data to be analyzed. One technique is to develop an algorithm specifically for large datasets. Koufakou and Georgiopoulos (2010) introduced a (distributed) two-step technique tailored to processing large and heterogeneous datasets. They use the categorical attributes to calculate an outlier score and partition the data and then use these partitions to analyze the continuous attributes and identify the outliers. Another approach is to keep the outlier detector simple. Eiweck et al. (2010) used simple statistical outlier detectors based on inter-quartile range and standard deviations to allow near real-time analysis of telecoms networks through analyzing the data using different time frames. The simplicity allows large data to be processed in near-real time. Bohm. Haegler, Müller, and Plant, (2009) aimed to be parameter free precluding the need to tune parameters and allowing fast training. They defined the data by a mixture model of distributions and a point is an outlier if it does not fit in any of the distribution functions.

Processing can also be speeded by only storing the required granularity of information. Businesses can store all of the raw data, samples of the data, summaries of the data or a combination, for example, storing a data summary locally for fast access and maintaining the raw data in large data warehouse to be accessed when necessary. Dash and Ng (2010) used sampling to analyze transactional data (supermarket sales) for outliers. The authors selected a representative sample from the entire dataset by ensuring that the distance between the sample and the dataset is below a threshold. This sample models normality so new transactions can be classified as outliers or normal by comparing them against the sample. However data sampling has to be performed carefully as spikes and outliers may be discarded during sampling as they occur infrequently and modelling normality requires clean data to ensure that no outliers are contaminating the training sample. Omniture (http://www.omniture.com/en/) and other commercial tools use data summarization. Storing a data summary allows a business to only save selected features of the data; vastly reducing the storage requirement and speeding processing. Also, some outliers can only be found in high dimensional data using feature subsets which represent smaller views of the data (Koufakou & Georgiopoulos, 2010). Again, summarization has to be performed carefully to ensure that outliers are not removed during feature selection.

Big Data warehouses store vast timelines of data for analysis. These data are often distributed across compute clusters or geographical sites. Data warehouse infrastructures have two main components: software to distribute and store the data for instance Hadoop Distributed File System (http://hadoop.apache.org/) and software to re-

Figure 2. The diagram on the left demonstrates local outlier detection where the outliers are determined locally and merged at the sink node. In contrast, global outlier detection merges the data at the sink node prior to outlier detection as shown on the right.



trieve and perform computations on the data from these distributed machines for instance MapReduce. To process large data, authors have taken existing outlier detection algorithms such as the binary KD-Tree (He, Ma, Wang, Zhuang, & Shi, 2011) or graphs representing social networks or Web page links (Kang, Chau, & Faloutsos, 2011) and parallelized them for distributed processing using Hadoop. Distributed processing can be *local* or *global*.

Outliers can be identified *locally* and the local outlier scores aggregated to produce a global outlier score at a coordinating node or the local data can be aggregated at a coordinating node and analyzed for outliers *globally*. The communication overhead is an important consideration for Big Data. Local outlier detection is likely to have the lowest overhead as only outlier scores need to be aggregated whereas a global approach needs to combine all data. However, a global approach can provide a system-wide view whereas local processing limits the overview. Anguilli, Basta, Lodi, and Sartori (in press) adapted a distance-based approach to detect outliers locally in very large datasets using a small subset of the dataset, computing distances on local nodes and iteratively merging the results at a coordinating node. They developed a ”lazy” version, which only sends distances when they are needed and this showed the most promising performance. Gao (2011) proposed to detect outliers globally by aggregating all different data sources into a matrix representation which preserves the individual object relationships. Computing cosine distance between the components of the eigenvectors of the matrix using spectral techniques can identify the key features of the combined matrix and pinpoint outliers. Das, Bhaduri, and Votava (2011) performed both local and global analyses by using a one-class SVM to identify outliers locally and collecting these local outliers at one coordinating node. The coordinating node also uses data samples from the local nodes to build a global model. All local outliers are tested against this global model.

When analyzing Big Data for outliers, the high number of outliers detected may overwhelm the system and some of these outliers detected may be false alarms. One option is to rank the outliers to assign a priority so the outliers can be analyzed in priority order as resources become available. Viswanathan et al. (2012) proposed a lightweight method for outlier ranking capable

of operating in modern data centers by calculating the outlier probability using simple statistics. Another approach is to validate anomalies once they have been detected to remove false positives. Sithirasenan and Muthukkumarasamy (2011) analyzed network intrusion data from different viewpoints and calculated an outlier score by comparing each outlier found with the data of its nearest neighbors using entropy.

Many authors have examined hybrid systems. As computer networks become ever more complex with grid or cloud systems, the task of network intrusion detection becomes ever more complicated. It needs to aggregate and analyze the data across multiple nodes and layers; process ever increasing volumes of data and detect ever more sophisticated attacks including coordinated and distributed attacks. Roschke, Cheng, and Meinel (2009) and Vieira et al. (2010) among others have developed hybrid modular intrusion detection systems to identify malicious behavior. Modules include a pattern-based anomaly detector to learn normal behavior and recognize novel (outlier) behavior; a rule-based signature detector which can be sophisticated but cannot detect novel attacks and frequency detectors to check for repeated behavior in a short space of time. Enterprise software running on computer networks is also becoming more complex spanning multiple computers, operating systems, languages and sites yet businesses need to ensure the availability and performance of enterprise software on their networks. Cherkasova, Ozonat, Mi, Symons, and Smirni (2009) proposed a hybrid system comprising two techniques: a regression-based model that learns the application's resource usage pattern and detects changes; and a performance signature that models the application's runtime behavior and identifies the causes of changes. Schlechtingen and Santos (2011) recommended hybrid techniques for different aspects of wind turbine monitoring. For example, they used regression to monitor simple components and neural networks for more complex monitoring tasks. Merdan et al. (2011) designed a system for factory process monitoring using communicating multi-agent systems. The agents analyze sensor data for anomalies locally and then communicate to generate a global view.

The financial sector has been proactive in using Big Data and business analytics for a wide range of tasks including fraud detection (Phua et al., 2010) and transaction processing (Dash & Ng, 2012). Recent research for financial fraud detection has examined using multi-layered approaches and harvesting data from social networks. Phua et al. (2012) developed a multi-layered system to identify anomalous credit applications in real-time. Credit application fraud frequently involves identity theft where the applicant purports to be someone else so their technique used real social relationships to generate a suspicion score. It also identified duplicate credit applications and increases the suspicion score. Šubelj et al. (2011) also analyzed social networks to find networks of collaborating fraudsters. They used an expert system to analyze the network across multiple layers. Konijn and Kowalczyk (2011) have investigated a hierarchical approach for detecting health insurance fraud which uses standard distance-based and density-based techniques to score outliers. It then aggregates these scores with a range of statistics calculated over different sections of the data to identify outliers in entities that are higher in the hierarchy. Other insurance fraud research has used survival analysis which analyses the time until a certain event and determines the probability that an individual will survive until a specified time. The individual could be an insurance policy owner and the policy survives while no fraudulent claims are made. Gepp, Kumar, Wilson, and Bhattacharya (2012) state that survival analysis is relatively new to business applications and will be assisted by Big Data which can provide historical timelines of data.

The transportation domain has been similarly proactive in incorporating Big Data and analyzing global views using distributed processing techniques. Accelerometer and GPS readings from mobile phones can act as probes to determine up-to-date road surface conditions and provide information to travelers and maintenance companies (Pertunnen et al., 2011). Road surface problems

manifest as anomalous patterns in the data. Other authors monitor distributed sensor data. Etienne, Devogele, and Bouju (2010) detected maritime traffic anomalies through distributed sensor monitoring; finding unusual trajectories of vessels using spatio-temporal pattern analysis of the sensor data. Das et al. (2010) analyzed sensor data from aircraft fleets using an adapted one-class multi-kernel SVM where different feature subsets train the different kernels. These methods show promise but there are still outstanding issues to allow real-time monitoring in these big transportation datasets including: distinguishing the individual objects in dense traffic; distinguishing the source of mobile phone signals, for example, pedestrians versus vehicles; and, having sufficiently accurate GPS locations to pinpoint which road the signal originates from where the road network is dense. Das et al. (2010) note that analyzing data across vehicle fleets is challenging as the data is large, complex, often heterogeneous and requires a system-level analysis.

FUTURE RESEARCH DIRECTIONS

While computational capability has increased massively in recent years, gaps still exist especially in capacity and speed for processing Big Data. However, High Performance Computing should progress to fill many of the gaps. Murphy (2011) pinpoints cloud-based data analytics as a rich area of future research and development with anomaly detection and other performance monitoring tools available as "as-a-service" tools. Such systems can use the large volumes of data for data mining and identify new patterns and trends. Thus, Big Data will produce a shift in the analysis from hypothesis-driven discoveries to data-driven discovery of patterns or anomalies in the data (Sanfilippo, Wolf, O'Connell, Carey, & Longstaff, 2012).

As the volume and richness of data expands then more outliers will be found. It will not be sufficient to just detect outliers; systems will need to provide precise and systematic explanations about the outliers to guide system analysts. Current explanations are often cumbersome and the relationships obtained are too complex to understand. As the number of anomalies detected increases then anomalies will also need to be prioritized and scheduled for investigation (Sanfilippo et al., 2012). Machine faults, hack attacks or traffic incident are critical anomalies that require immediate detection and investigation. Other anomalies are more strategic and can be scheduled for less busy processing periods. Additionally, detected anomalies can be screened to remove false positives and reduce the number of anomaly investigations. Outlier detection will also need to consider the cost of misclassification. False negatives are usually more costly (such as missing instances of fraud or the precursors to an industrial process failure) than false positive errors which just waste the analyst's time investigating false leads.

Existing outlier algorithms can be adapted for Big Data or new techniques considered. A potential new technique is Deep Learning (Bengio, 2009) which generates data models comprising multiple levels of non-linear operations and is well suited to the complexity challenges of Big Data. There is also a move to integrate different outlier algorithms to analyze Big Data. Hybrid techniques should give better performance than the individual algorithms by overcoming their individual limitations and exploiting their different strengths. The use of hybrid techniques will expand as the size and range of the data and the variety of outliers expands.

CONCLUSION

Massive and complex data sources appear at first glance to be part of the challenge for data mining tasks such as outlier detection but they also hold many opportunities. New outliers can be uncovered, vast timelines of data are available for analysis and the data models learned will be

increasingly rich as the data expands. This will allow businesses to learn more about market trends, economic factors, competitors and customers. Existing outlier detection methods need to be adapted or new methods devised to process the complexity, volume, speed and variety of Big Data. Also, how the data is represented needs to be carefully considered including using data samples and feature subsets. Big Data is often distributed and streamed so outlier detection needs to be able to process distributed and streaming data sources. Hybrid outlier detectors are often used for Big Data to exploit the power of the individual methods. The sheer volume of outliers detected in Big Data will necessitate screening and/or prioritizing so outliers can be investigated systematically and systematic explanations provided regarding why a data point is an outlier.

The future is likely to see outlier detection available on high performance computers as software-as-a-service applications.

Outlier detection on Big Data will uncover previously unseen outliers, fill in gaps in outlier knowledge, provide new insights and identify new data relationships.

REFERENCES

Angiulli, F., Basta, S., Lodi, S., & Sartori, C. (in press). Distributed strategies for mining outliers in large data sets. *IEEE Transactions on Knowledge and Data Engineering*.

Bengio, Y. (2009). Learning deep architectures for AI. *Foundation and Trends in Machine Learning*, *2*(1), 1–127. doi:10.1561/2200000006

Böhm, C., Haegler, K., Müller, N., & Plant, C. (2009). CoCo: coding cost for parameter-free outlier detection. In, *Proceedings of 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 149-158). New York, USA: ACM.

Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. *ACM Computing Surveys*, *41*(3). doi:10.1145/1541880.1541882

Chaudhary, K., Yadav, J., & Mallick, B. (2012). A review of fraud detection techniques: Credit card. *International Journal of Computers and Applications*, *45*(1), 39–44.

Cherkasova, L., Ozonat, K., Mi, N., Symons, J., & Smirni, E. (2009). Automated anomaly detection and performance modeling of enterprise applications. *ACM Transactions on Computer Systems*, *27*(3). doi:10.1145/1629087.1629089

Das, K., Bhaduri, K., & Votava, P. (2011). Distributed anomaly detection using 1-class SVM for vertically partitioned data. *Statistical Analysis and Data Mining*, *4*(4), 393–406. doi:10.1002/sam.10125

Das, S., Srivastava, A., Matthews, B., & Oza, N. (2010). Multiple kernel learning for heterogeneous anomaly detection: Algorithm and aviation safety case study. In, *Proceedings of 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 47-56). New York, USA: ACM.

Dash, M., & Ng, W. (2010). Outlier detection in transactional data. *Intelligent Data Analysis*, *14*(3), 283–298.

Eiweck, J., Pattinson, C., Behringer, R., & Seewald, A. (2010). A new approach for outlier detection in near real-time. In, *Proceedings of 4th UKSim European Symposium on Computer Modeling and Simulation* (pp. 477-483). Washington DC, USA: IEEE Computer Society.

Ertöz, L., Steinbach, M., & Kumar, V. (2003). Finding clusters of different sizes, shapes, and densities in noisy, high dimensional data. In, *Proceedings of 3rd SIAM International Conference on Data Mining* (pp. 47-58). Philadelphia, USA: Society for Industrial and Applied Mathematics.

Etienne, L., Devogele, T., & Bouju, A. (2010). Spatio-temporal trajectory analysis of mobile objects following the same itinerary. In, *Proceedings of the International Symposium on Spatial Data Handling (SDH)*, Hong Kong.

Fang, Z., Luo, G., Xu, S., & Fei, F. (2009). Stock fluctuations anomaly detection based on wavelet modulus maxima. In, *Proceedings of 2nd International Conference on Business Intelligence and Financial Engineering* (pp. 360-363). Washington DC, USA: IEEE Computer Society.

Gao, J. (2011). *Exploring the power of heterogeneous information sources*, Unpublished doctoral dissertation, University of Illinois at Urbana-Champaign.

Gepp, A., Kumar, K., Wilson, J., & Bhattacharya, S. (2012). A comparative analysis of decision trees vis-à-vis other computational data mining techniques in automotive insurance fraud detection. *Journal of Data Science*, *10*(3), 537–561.

He, Q., Ma, Y., Wang, Q., Zhuang, F., & Shi, Z. (2011). Parallel outlier detection using kd-tree based on MapReduce. In, *Proceedings of 3rd International Conference on Cloud Computing Technology and Science* (pp. 75-80). Washington DC, USA: IEEE Computer Society.

Hodge, V. (2011). *Outlier and anomaly detection: A survey of outlier and anomaly detection methods*. Saarbrucken, Germany: Lambert Academic Publishing.

Kang, U., Chau, D., & Faloutsos, C. (2011). Mining large graphs: Algorithms, inference, and discoveries. In, *Proceedings of 27th International Conference on Data Engineering* (pp. 243-254). Washington DC, USA: IEEE Computer Society.

Kimmel, R., Booth, D., & Booth, S. (2010). The analysis of outlying data points by robust locally weighted scatter plot smooth: a model for the identification of problem banks. *International Journal of Operational Research*, *7*(1), 1–15. doi:10.1504/IJOR.2010.029514

Konijn, R., & Kowalczyk, W. (2011). Finding fraud in health insurance data with two-layer outlier detection approach. In, *Proceedings of 13th International Conference on Data Warehousing and Knowledge Discovery* (pp. 394-405). Berlin, Germany: Springer-Verlag.

Koufakou, A., & Georgiopoulos, M. (2010). A fast outlier detection strategy for distributed high-dimensional data sets with mixed attributes. *Data Mining and Knowledge Discovery*, *20*(2), 259–289. doi:10.1007/s10618-009-0148-z

Merdan, M., Vallee, M., Lepuschitz, W., & Zoitl, A. (2011). Monitoring and diagnostics of industrial systems using automation agents. *International Journal of Production Research*, *49*(5), 1497–1509. doi:10.1080/00207543.2010.526368

Murphy, J. (2011). Performance engineering for cloud computing. In, *Proceedings of 8th European conference on Computer Performance Engineering* (pp. 1-9), Berlin, Germany: Springer-Verlag.

Perttunen, M., Mazhelis, O., Cong, F., Kauppila, M., Leppänen, T., Kantola, J., et al. (2011). Distributed road surface condition monitoring using mobile phones. In, *Proceedings of 8th International Conference on Ubiquitous Intelligence and Computing* (pp. 64-78). Berlin, Germany: Springer-Verlag.

Phua, C., Lee, V., Smith-Miles, K., & Gayler, R. (2010). A comprehensive survey of data mining-based fraud detection, *Research Computing Research Repository*, abs/1009.6119.

Phua, C., Smith-Miles, K., Lee, V., & Gayler, R. (2012). Resilient identity crime detection. *IEEE Transactions on Knowledge and Data Engineering*, *24*(3), 533–546. doi:10.1109/TKDE.2010.262

Roschke, S., Cheng, F., & Meinel, C. (2009). Intrusion detection in the cloud. In, *8th IEEE International Conference on Dependable, Autonomic and Secure Computing* (pp. 729-734). Washington DC, USA: IEEE Computer Society.

Sanfilippo, A., Wolf, D., O'Connell, K., Carey, L., & Longstaff, T. (2012) *C3E Workshop 2011 final report*. Retrieved July 09, 2012 from http://cps-vo.org/node/3456.

Schlechtingen, M., & Santos, I. (2011). Comparative analysis of neural network and regression based condition monitoring approaches for wind turbine fault detection. *Mechanical Systems and Signal Processing*, *25*(5), 1849–1875. doi:10.1016/j.ymssp.2010.12.007

Sithirasenan, E., & Muthukkumarasamy, V. (2011). Substantiating anomalies in wireless networks using group outlier scores. *Journal of Software*, *6*(4), 678–689. doi:10.4304/jsw.6.4.678-689

Šubelj, L., Furlan, Š., & Bajec, M. (2011). An expert system for detecting automobile insurance fraud using social network analysis. *Expert Systems with Applications*, *38*(1), 1039–1052. doi:10.1016/j.eswa.2010.07.143

Vieira, K., Schulter, A., Westphall, C. B., & Westphall, C. M. (2010). Intrusion detection for grid and cloud computing. *IT Professional*, *12*(4), 38–43. doi:10.1109/MITP.2009.89

Viswanathan, K., Choudur, L., Talwar, V., Wang, C., MacDonald, G., & Satterfield, W. (2012). Ranking anomalies in data centers. In, *Proceedings of 13th IEEE/IFIP Network Operations and Management Symposium* (pp. 79-87). IEEE.

Yoon, J., & Kim, K. (2012). Detecting signals of new technological opportunities using SAO-based semantic patent analysis and outlier detection. *Scientometrics*, *90*(2), 445–461. doi:10.1007/s11192-011-0543-2

ADDITIONAL READING

Aggarwal, C., Zhao, Y., & Yu, P. (2011). Outlier detection in graph streams. In, *Proceedings IEEE 27th International Conference on Data Engineering* (pp. 399-409). Washington DC, USA: IEEE Computer Society.

Barnett, V., & Lewis, T. (1994). *Outliers in statistical data* (3rd ed.). Chichester, UK: John Wiley & Sons Ltd.

Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. *ACM Computing Surveys*, *41*(3). doi:10.1145/1541880.1541882

Franks, B. (2012). *Taming the big data tidal wave: Finding opportunities in huge data streams with advanced analytics*. Hoboken, NJ: John Wiley & Sons.

Gogoi, P., Bhattacharyya, D., Borah, B., & Kalita, J. (2011). A survey of outlier detection methods in network anomaly identification. *The Computer Journal*, *54*(4), 570–588. doi:10.1093/comjnl/bxr026

Hadi, A., Imon, A., & Werner, M. (2009). Detection of outliers. *Wiley Interdisciplinary Reviews: Computational Statistics*, *1*(1), 57–70. doi:10.1002/wics.6

Han, J., Kamber, M., & Pei, J. (2011). *Data mining: Concepts and techniques* (3rd ed.). Waltham, MA: Morgan Kaufmann Publishers.

Hawkins, D. (1980). *Identification of Outliers*. London, UK: Chapman and Hall. doi:10.1007/978-94-015-3994-4

Hodge, V. (2011). *Outlier and anomaly detection: A survey of outlier and anomaly detection methods*. Saarbrucken, Germany: Lambert Academic Publishing.

Hodge, V., & Austin, J. (2004). A survey of outlier detection methodologies. *Artificial Intelligence Review*, *22*(2), 85–126. doi:10.1023/B:AIRE.0000045502.10941.a9

Patcha, A., & Park, J. (2007). An overview of anomaly detection techniques: Existing solutions and latest technological trends. *Computer Networks*, *51*(12), 3448–3470. doi:10.1016/j.comnet.2007.02.001

Phua, C., Lee, V., Smith-Miles, K., & Gayler, R. (2010). A comprehensive survey of data mining-based fraud detection, *Research Computing Research Repository*, abs/1009.6119.

Su, X., & Tsai, C.-L. (2011). Outlier detection. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *1*(3), 261–268.

Wang, J. (2009). *Encyclopedia of data warehousing and mining* (2nd ed.). Hershey, PA: IGI Global.

Yang, P., & Zhu, Q. (2011). Finding key attribute subset in dataset for outlier detection. *Knowledge-Based Systems*, *24*(2), 269–274. doi:10.1016/j.knosys.2010.09.003

Zhang, Y., Meratnia, N., & Havinga, P. (2010). Outlier detection techniques for wireless sensor networks: A survey. *IEEE Communications Surveys & Tutorials*, *12*(2), 159–170. doi:10.1109/SURV.2010.021510.00088

KEY TERMS AND DEFINITIONS

Anomaly Detection: The task of finding anomalies in a business's data. Some authors use "anomaly detection" to specifically refer to network intrusion detection.

Anomaly: Datum that deviates from the norm (often used interchangeably with "outlier").

Big Data: Large, dynamic and unstructured collections of data often distributed and streamed.

Business Analytics: The analysis of a business's data to gain insight into the business.

Data Mining: The process of analyzing data from different perspectives to predict future behavior and trends.

Distributed: Data storage and processing that is performed in different locations connected by transmission links.

Fault Detection: The task of finding failures in hardware or software.

Outlier Detection: The task of finding outliers in a business's data. It is considered a fundamental task in data mining.

Outlier: Datum that deviates from the norm.