Outlier Detection in Big Data

Victoria J. Hodge
University of York, UK

INTRODUCTION

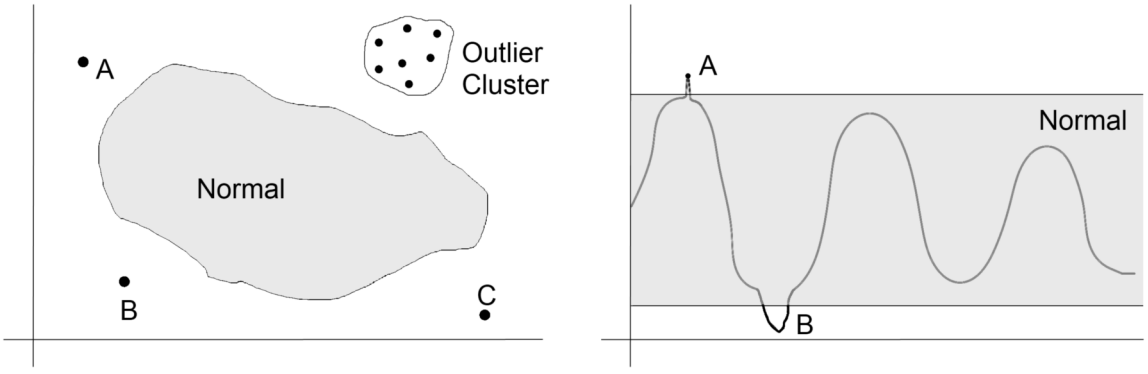
This chapter will examine the issues posed by Big Data for the task of outlier detection. An outlier (Hodge, 2011) (often called an anomaly (Chandola, Banerjee, & Kumar, 2009) in the literature) is a particular data point or, in some instances, a small set of data points that is inconsistent with the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