Anomaly Detection in Finance: Editors' Introduction

Archana Anandakrishnan ARCHANA.ANANDAKRISHNAN@AEXP.COM
American Express

Senthil Kumar SENTHIL.KUMAR@CAPITALONE.COM
Capital One

Alexander Statnikov ALEXANDER.STATNIKOV@AEXP.COM
American Express

Tanveer Faruquie TANVEER.FARUQUIE@CAPITALONE.COM
Capital One

Di Xu DI.W.XU@AEXP.COM
American Express

1. Introduction

Detecting anomalies and novel events is vital to the financial services industry. These events may often be indicative of illegal activities such as fraud, risk, identity theft, network intrusion, account takeover and money laundering which may result in undesired outcomes such as disruption in service and other breakdowns. As financial environments change, digital adoption grows and the data moves at increasing speed and volume, the problem of detecting anomalies in real time at large scale becomes increasingly challenging. This is further compounded by the fact that more and more anomaly detection applications require operational decision making in real time. Several new ideas are emerging to tackle this challenge, including semi-supervised learning methods, deep learning based approaches and network/graph based solutions. These approaches must often be able to work in real time by consuming and processing large volumes of data produced in real time. The 2017 KDD Workshop on Anomaly Detection in Finance held at Halifax, Nova Scotia on Aug 14, 2017, brought together researchers and practitioners to discuss these new approaches and solutions. This half-day workshop consisted of two keynote speeches, two industry perspective talks, five full papers and eight spotlight talks. The talks and papers covered many ideas that are of general interest to the entire anomaly detection community in addition to discussing challenges specific to the financial services industry. We are happy to publish a select set of papers from the workshop in this issue of the Proceedings of Machine Learning Research.

We open this Volume with a brief introduction to the applications in the financial services industry and how the papers in this Volume address the challenges for these applications.

2. Applications in the Financial Services Industry

Multiple industries are witnessing an exponential increase in the availability of streaming large volume of data. This is certainly true for the financial industry. Largely driven by an increase in instrumentation both on the front end and back end applications, we now have enormous number of systems that produce continuously changing data in real time. This data is representative of the health and well-being of the system and applications. Hence, any deviations from the past behavior that is unusual and significant is of special interest and may require action. Detecting anomalies is thus increasingly becoming the core of many business operations.

An anomaly is an observation which deviates so much from the other observations as to create suspicion that it was generated by a different mechanism (Hawkins, 1980). Anomalies are often indicative of something interesting, such as an unusually high demand, or something gone wrong, such as imminent failure and, as such, anomaly detection has received considerable attention in many different domains. Anomaly detection techniques are generally based on determining what constitutes normal data and identifying deviations from the normal. The problem of anomaly detection is difficult for a variety of reasons. First, it is often difficult to define what the normal observations are. Second, the normal data can change with time. For example, a customer purchasing behavior can evolve over time and it is easy to mistake the purchase of an entirely new item as an anomaly. It is not just the normal data that might evolve over time, but in areas such as cyber security and credit card fraud, adversaries keep adapting their attacks and hence the nature of the anomalies can vary with time. A good exposition to different types of anomalies and detection methods is presented in Chandola et al. (2009).

Depending on the availability of training data, anomaly detection techniques naturally fall into two categories, supervised and unsupervised. However, even when training data is available, the fraction of anomalous observations is usually very small, and we must be careful to take this class-imbalance into account. In most anomaly detection applications, there is no labeled training data available and thus it is no surprise that majority of anomaly detection algorithms are unsupervised. There might be special situations where we have examples of normal observations but no examples of anomalies, or where there is a very limited set of labeled data for anomalous and normal observations and plenty of unlabeled data. Comparison of some of these techniques is provided in Emmott et al. (2013).

Many of these methods are applied in different situations arising in the financial industry, especially the credit card industry. Some of the popular use cases of anomaly detection techniques in the financial services industry include Phua et al. (2010); Paula et al. (2016); Wang et al. (2017):

1. Transactional Fraud
2. Anti-Money Laundering
3. Identity Theft and Fake Account Registration
4. Risk Modeling
5. Account Takeover

6. Promotion Abuse
7. Customer Behavior Analytics
8. Cyber Security

These applications necessitate the use of Machine Learning algorithms and methods. The nature and the volume of the data constantly drives practitioners to seek out new and improved methods that can efficiently help deliver solutions to these problems. The papers presented at this workshop and subsequently in this Volume discuss advancements to these broad themes: innovative approaches and novel applications.

3. Overview of Contributions to this Volume

Many of the ideas presented here are of general interest to the entire anomaly detection community and are not unique to the financial industry. We list some of the broad challenges that the papers in this Volume address and how these could be extra beneficial to the financial services industry.

3.1. Lack of labeled data

If is often extremely difficult to find or prepare data that contains labels to indicate anomalies. Even if the labels exist for some classes of anomalies, events such as fraud and money laundering are rapidly evolving and there are many scenarios where it is not possible to create labels. Unsupervised methods are a natural choice and have been popular for anomaly detection (Guthrie et al., 2007; Goldstein and Uchida, 2016). We start the Volume with the keynote address that is a broad survey of a range of methods to Uncover Unknown Unknowns in data by using unsupervised methodologies (Shabat et al., 2017). In addition, the authors also propose new methods in this class for anomaly detection. The works presented by Ram and Gray (2017) and Kashef (2017) describe enhancements to existing unsupervised algorithms that could make them more superior. The authors Ram and Gray (2017) propose Density Estimation Trees, an adaptation of decision trees in unsupervised settings and similarly Kashef (2017) proposes cooperative agreement from an ensemble of clustering algorithms respectively as novel unsupervised methods to detect anomalies. Detection of anomalies often relies on time series data and looks for patterns that seem out of the ordinary based on past behavior. This setting typically may or may not have explicitly labelled data and often uses the past data as a source of normal behavior. The authors van Adelsberg and Schwantes (2017) discuss some real challenges with this approach around the need of vast amount of past data and present a method for learning time series behavior on multiple timescales with the use of a binned process that could lead to improved forecasts and lower false positive alert rates. Yet another powerful approach to mining large datasets for insights is to build networks and graphs that would connect similar entities (Akoglu et al., 2015; Noble and Cook, 2003). The advent of very powerful in-memory computation platforms like Spark make it a real possibility to use Graph based methods for real applications. The authors Cao et al. (2017) present a system to detect fraud by capturing common fraudulent behavior in networks.

3.2. Signal to noise ratio

Class imbalance is a relevant problem for anomaly detection. Normal data is present in large quantities and the anomalous instances are usually very small in number. The classic example of this is again transaction fraud, which is fortunately far fewer in number as compared to the number of legitimate transactions carried out by millions of customers every day. This makes anomaly detection akin to finding the needle in a haystack. Intelligent feature engineering, sampling and preprocessing methods can alleviate this problem (Bahnsen et al., 2016; Davis and Clark, 2011; Lee and Stolfo, 2000). The second paper in this Volume (Ye, 2017), also the second keynote at the workshop presents three analytical techniques that can help in the efficient separation of signal from noise.

3.3. Novel Applications

The papers by Miller and Cezeaux (2017) and Lasaga and Santhana (2017) demonstrate novel applications of known anomaly detection methods to real problems, providing great benchmarks to the scalability and performance of these algorithms. For instance, Miller and Cezeaux (2017) use isolation forest algorithm for the detection of agents correlated to adverse outcomes. Similarly, the authors Lasaga and Santhana (2017) use Deep Learning to identify treatment fraud amongst Healthcare Providers detecting unnecessary treatment procedures and prescriptions.

3.4. Automation

Novel applications as described in Sec. 3.3 need automation and in many cases real-time implementations to be successful and revenue generating. Two expositions of how companies are taking these methods to the next level by implementing machine learning into production is discussed by Toledano et al. (2017) and Aggarwal et al. (2017). Toledano et al. (2017) discuss how a real-time anomaly detection system chooses between algorithms for implementation and the architecture used by the implementation. This system is reported to be adopted by 50 different companies reporting over 120 million time series metrics constantly. Aggarwal et al. (2017) discuss an automated approach to identify data attribute anomalies for the purpose of improving data quality. This ensures that models and decision support systems do not intake incorrect data. This is a demonstration of anomaly detection of a very different type. It is not a case where anomalies are necessarily malicious by intention, but are anomalies created by human error or system errors and if left unhandled, could lead to subpar data products and applications.

3.5. Interpretability

Interpretability of models and results is increasingly becoming a key to the adoption of Machine Learning, especially in financial services industries (Vellido et al., 2012). This allows users to infer actionable insights from the results. The importance of interpretable models is discussed by the authors of multiple papers in this Volume. For instance, Ram and Gray (2017) (density estimation trees) pay special attention to categorical variables and refrain from techniques such as one-hot encoding for categorical variables to preserve interpretability. Redefining the data in terms of events on a time series allows for the use of

algorithms such as, frequent pattern mining which can be used to identify the occurrence of exact events that predict a record to be anomaly. The authors Kuchar and Svatek (2017) propose combining the Frequent Pattern Mining with Isolation Forest to define a robust anomaly score. Finally, the use of interactive visualization of data also helps in the quick interpretation of the results as presented in Aggarwal et al. (2017).

3.6. Definition of New Features

A time tested strategy of extracting new information from data is to define new features from human intuition that is not easily learned by simple models. Long before the advent of complex methods such as ensemble based approaches or deep neural networks, modelers have fed in complex logics to their simple models and have obtained phenomenal results. The advantage of such an approach is the fact that they appeal to the human intuition and make it easy to understand the underlying dynamics of a system. Such methods are not only common in finance, but in almost all fields of science. The authors Ki and Yoon (2017) propose new features based on purchase density that could be beneficial for fraud detection systems even when the system has not seen the transaction behavior of a specific customer.

The works presented here are by no means an exhaustive coverage of anomaly detection methods for finance. They are however an indication of the current status of a field that is under rapid progress. The studies presented highlight a few critical learnings: First, there is no clear singular algorithm that can be called the "answer" to anomaly detection in finance. The key is in clever combinations of existing methods and pipelining them intelligently. Second, users looking for benchmarks on how popular algorithms can scale on real applications can turn to the papers in this Volume. Some of the applications by the industry speakers present real use cases and results that are hard to find elsewhere in literature. Finally, the works also highlight the key challenges that are currently being solved in this industry and make it clear that there is plenty of opportunity for innovative ideas around anomaly detection.

4. Acknowledgements

These proceedings contain 14 contributed papers that were presented at the Anomaly Detection in Finance Workshop at KDD, held in Halifax, Nova Scotia, Canada on August 14, 2017. We thank the keynote speakers and participants who enthusiastically submitted and presented their work, and the Program Committee who helped with the organization of this conference.

Organizers

Program Committee

Ccile Germain (Universit Paris-Sud, France), Yinglian Xie (DataVisor), Khalid Benabdeslem (University of Lyon), Stephen Clemencon (Telecom-ParisTech), Ignacio Arnaldo Lucas (PaternEx / MIT), Danny Silver (Acadia University)

References

Leman Akoglu, Hanghang Tong, and Danai Koutra. Graph based anomaly detection and description: a survey. *Data Mining and Knowledge Discovery*, 29(3):626–688, 2015.

Alejandro Correa Bahnsen, Djamila Aouada, Aleksandar Stojanovic, and Björn Ottersten. Feature engineering strategies for credit card fraud detection. *Expert Systems With Applications*, 51:134–142, 2016.

Bokai Cao, Mia Mao, Siim Viidu, and Philip Yu. Collective fraud detection capturing inter-transaction dependency. pages 66–75, 2017.

V. Chandola, A. Banerjee, and V. Kumar. Anomaly detection: A survey. *CM Computing Surveys*, 41(3), 2009.

Jonathan J Davis and Andrew J Clark. Data preprocessing for anomaly based network intrusion detection: A review. *Computers & Security*, 30(6):353–375, 2011.

Andrew F. Emmott, Shubhomoy Das, Thomas Dietterich, Alan Fern, and Weng-Keen Wong. Systematic construction of anomaly detection benchmarks from real data. In *Proceedings of the ACM SIGKDD Workshop on Outlier Detection and Description*, ODD '13, pages 16–21, New York, NY, USA, 2013. ACM. ISBN 978-1-4503-2335-2. doi: 10.1145/2500853.2500858. URL http://doi.acm.org/10.1145/2500853.2500858.

Markus Goldstein and Seiichi Uchida. A comparative evaluation of unsupervised anomaly detection algorithms for multivariate data. *PloS one*, 11(4):e0152173, 2016.

David Guthrie, Louise Guthrie, Ben Allison, and Yorick Wilks. Unsupervised anomaly detection. In *IJCAI*, pages 1624–1628, 2007.

Douglas M Hawkins. *Identification of outliers*, volume 11. Springer, 1980.

R.F. Kashef. Ensemble-based anomaly detetction using cooperative learning. pages 43–55, 2017.

Youngjoon Ki and Ji Won Yoon. Pd-fds: Purchase density based online credit card fraud detection system. pages 76–84, 2017.

Jaroslav Kuchar and Vojtech Svatek. Spotlighting anomalies using frequent patterns. pages 33–42, 2017.

Daniel Lasaga and Prakash Santhana. Deep learning to detect medical treatment fraud. pages 114–120, 2017.

Wenke Lee and Salvatore J Stolfo. A framework for constructing features and models for intrusion detection systems. *ACM transactions on Information and system security (TiSSEC)*, 3(4):227–261, 2000.

Michelle Miller and Robert Cezeaux. Sleuthing for adverse outcomes: Using anomaly detection to identify unusual behaviors of third-party agents. pages 121–125, 2017.

Caleb C Noble and Diane J Cook. Graph-based anomaly detection. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 631–636. ACM, 2003.

Ebberth L Paula, Marcelo Ladeira, Rommel N Carvalho, and Thiago Marzagão. Deep learning anomaly detection as support fraud investigation in brazilian exports and anti-money laundering. In *Machine Learning and Applications (ICMLA), 2016 15th IEEE International Conference on*, pages 954–960. IEEE, 2016.

Clifton Phua, Vincent Lee, Kate Smith, and Ross Gayler. A comprehensive survey of data mining-based fraud detection research. *arXiv preprint arXiv:1009.6119*, 2010.

Parikshit Ram and Alexander G. Gray. Fraud detection with density estimation trees. pages 85–94, 2017.

Gil Shabat, David Segev, and Amir Averbuch. Uncovering unknown unknowns in financial services big data by unsupervised methodologies: Present and future trends. pages 8–19, 2017.

Meir Toledano, Ira Cohen, Yonatan Ben-Simhon, and Inbal Tadeski. Real-time anomaly detection system for time series at scale. pages 56–65, 2017.

Matthew van Adelsberg and Christian Schwantes. Binned kernels for anomaly detection in multi-timescale data using gaussian processes. pages 102–113, 2017.

Alfredo Vellido, José David Martín-Guerrero, and Paulo JG Lisboa. Making machine learning models interpretable. In *ESANN*, volume 12, pages 163–172, 2012.

Cheng Wang, Bo Yang, and Jing Luo. Identity theft detection in mobile social networks using behavioral semantics. In *Smart Computing (SMARTCOMP), 2017 IEEE International Conference on*, pages 1–3. IEEE, 2017.

Nong Ye. Analytical techniques for anomaly detection through features, signal-noise separation and partial-value association. pages 20–32, 2017.