Big Data Classification: Problems and Challenges in Network Intrusion Prediction with Machine Learning

Shan Suthaharan
Department of Computer Science
University of North Carolina at Greensboro,
Greensboro, NC 27402, USA
+1 336 256 1122
s_suthah@uncg.edu

ABSTRACT

This paper focuses on the specific problem of Big Data classification of network intrusion traffic. It discusses the system challenges presented by the Big Data problems associated with network intrusion prediction. The prediction of a possible intrusion attack in a network requires continuous collection of traffic data and learning of their characteristics on the fly. The continuous collection of traffic data by the network leads to Big Data problems that are caused by the volume, variety and velocity properties of Big Data. The learning of the network characteristics requires machine learning techniques that capture global knowledge of the traffic patterns. The Big Data properties will lead to significant system challenges to implement machine learning frameworks. This paper discusses the problems and challenges in handling Big Data classification using geometric representation-learning techniques and the modern Big Data networking technologies. In particular this paper discusses the issues related to combining supervised learning techniques, representation-learning techniques, machine lifelong learning techniques and Big Data technologies (e.g. Hadoop, Hive and Cloud) for solving network traffic classification problems.

Categories and Subject Descriptors

C.2.4 [**Distributed Systems**]: Distributed applications.

General Terms

Measurement, Performance, Experimentation, Design and Security

Keywords

Big Data, Hadoop distributed file systems, intrusion detection, machine learning

1. INTRODUCTION

Big Data is currently defined using three data characteristics: volume, variety and velocity [1]. It means that some point in time, when the volume, variety and velocity of the data are increased, the current techniques and technologies may not be able to handle storage and processing of the data. At that point the data is defined as Big Data. In the Big Data research, the term Big Data Analytics is defined as the process of analyzing and understanding the characteristics of massive size datasets by extracting useful geometric and statistical patterns. Ideally these three characteristics of a dataset increase the complexity of the data and thus make the current techniques and technologies stop functioning as expected within a given processing time. Many applications suffer from the Big Data problem, including network traffic risk analysis, geospatial classification and business forecasting. Network intrusion detection and prediction are time sensitive applications and they require highly efficient Big Data techniques and technologies to tackle the problem on the fly. The new technologies can help conduct Big Data analytics on various applications. The techniques, Hadoop Distributed File Systems (HDFS) [2], Cloud technology [3] and Hive database [2] can be combined to address the problems like Big Data classification. However the applications that require continuous growth in the Big Data domain, including intrusion prediction system and geospatial can suffer from the Big Data problems significantly.

In this paper some of the problems and challenges associated with the integration of modern networking technologies and machine learning techniques for solving Big Data classification problem for network intrusion prediction are discussed.

2. PROOF OF BIG DATA

The first challenging problem rests on the current definition of Big Data; how to prove (or show) that the network traffic data satisfy the Big Data characteristics for Big Data classification. This is the first important task to address in order to make the Big Data analytics efficient and cost effective. The early detection of the Big Data characteristics can provide a cost effective strategy to many organizations to avoid unnecessary deployment of Big Data technologies. The data analytics on some data may not require Big Data techniques and technologies; the current and well established techniques and technologies maybe sufficient to handle the data storage and data processing. Hence we need an early analysis and understanding of the data characteristics for classification.

The current research in the field of Big Data has ignored the early detection of Big Data characteristics. For example the current definition of Big Data defined on a 3D space, V^3, formed by three parameters, volume, variety and velocity cannot provide a suitable platform for the early detection of Big Data characteristics for Big Data classification. Figure 1 shows the 3D space defined for Big Data, where the axis of volume represents the growth of data size, the axis of velocity represents the increase in speed in which the data must be processed, and the axis of variety represents the increase in various types of data. Suppose the dataset has *n* number of zeros, *n* number of ones, *n* number of twos and so on, and continuously growing to infinity, then the space V^3 will suggest it as Big Data, but a sampling will simply suggest it as small data. Therefore a deeper preliminary analysis is required to determine the Big Data characteristics. To alleviate this problem this paper suggests a new definition for Big Data by introducing a 3D space, C^3, which is defined based on three new parameters, cardinality, continuity, and complexity.

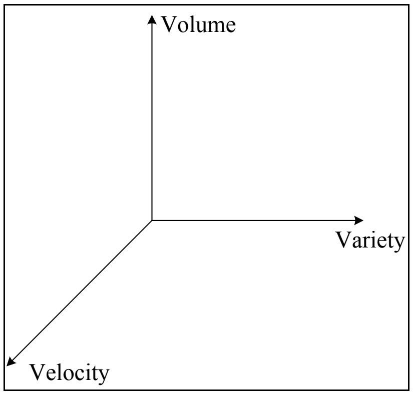

Figure 1: Current definition (V^3) of Big Data characteristics

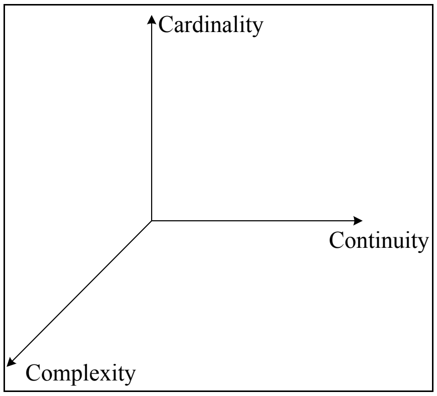

Figure 2: Proposed definition (C^3) of Big Data characteristics

Compared to defining a metric to measure the Big Data characteristics in V^3 space, it is much easier to develop a metric in C^3 space using mathematical and statistical tools. In C^3 space the cardinality defines the number of records in the dynamically growing dataset at a particular instance. The continuity defines two characteristics and they are: (i) representation of data by continuous functions, and (ii) continuously growth of data size with respect to time. The complexity defines three characteristics and they are: (i) large varieties of data types, (ii) high dimensional dataset; and (iii) the speed of data processing is very high.

3. MANAGEMENT OF BIG DATA

The cardinality parameter demands the need for an efficient distributed file system for data capture, storage and analysis of network traffic for intrusion prediction. Then the continuity and complexity parameters add extra difficulties to the task of managing the Big Data. Hence the network topology must be designed in such a way that the Big Data Analytics problem can be handled efficiently with cost effectiveness objectives.

3.1 Network Topology

The modern computer technologies, like HDFS and public cloud, can help alleviate the cardinality problem in Big Data analytics. They can be integrated to build a large and flexible network topology with a storage infrastructure that can change adaptively based on the need of the Big Data processing requirements. However this integrated model will bring several challenges that must be handled efficiently. One such model is presented in Figure 3. It consists of four units: user Interaction and Learning System (UILS), Network Traffic Recording System (NTRS), HDFS and Cloud Computing Storage System (CCSS).

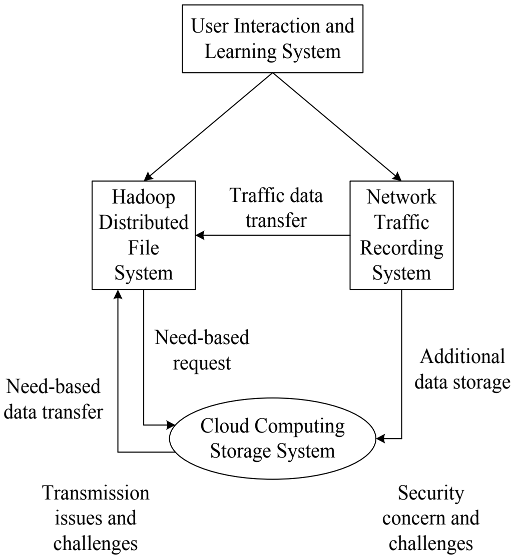
User Interaction and Learning System
Hadoop Distributed File System
Traffic data transfer
Network Traffic Recording System
Need-based request
Additional data storage
Need-based data transfer
Cloud Computing Storage System
Transmission issues and challenges
Security concern and challenges

Figure 3: Suggested network topology for Big Data analytics

The NTRS unit helps to capture network traffic and streams the traffic data to HDFS unit or CCSS unit in real-time based on the need of an additional storage. The HDFS system will also use Hive database to store the data. The UILS unit can learn and control the additional storage and data requirements.

3.2 Communication Challenges

In computer networking research and applications, the communication cost is the major concern compared to the processing cost of the data – the same in this suggested topology. The challenge here is to minimize that communication cost while satisfying the additional storage and data requirement from public cloud for processing Big Data. As stated in [3] and [4], the bandwidth and latency are the two major network features that will affect the communication between the clients and the cloud server. These problems and associated challenges to find solutions will adversely affect the timing requirements of the Big Data processing at HDFS and UILS. Hence the machine learning techniques, using these technologies, must be developed by keeping these problems and challenges in mind.

3.3 Security Challenges

The security mechanism in cloud technology is generally weak. Hence tampering of data at the public cloud is inevitable and it is a big concern. Finding a robust security mechanism for the purpose of using the public cloud like CCSS is a challenging problem. As stated in [5], in cloud technology, an attacker can easily tamper the data that is being exchanged between the CCSS server and the HDFS and NTRS units; the attacker can spoof the reply between them and shut down the server (CCSS) using DOS attack. These problems can lead to challenges in implementing Big Data analytics tool with the suggested network topology.

4. LEARNING OF BIG DATA

The complexity parameter consists of many other influential data characteristics, which are high dimensional dataset, a large number of data types (classes), high speed in which the data should be processed and unstructured data. The complexity parameter and its resulting problems have to be addressed using machine learning techniques. However the challenge still rests on improving the current learning techniques to deal with the Big Data classification problems and requirements.

4.1 Machine Learning

The traditional Machine Learning (ML) techniques have been developed and used for extracting useful information from the data through training and validation using labeled datasets. Three major problems that make the ML techniques unsuitable for solving Big data classification problems are: (1) An ML technique that is trained on a particular labeled datasets or data domain may not be suitable for another dataset or data domain – that the classification may not be robust over different datasets or data domains; (2) An ML technique is in general trained using a certain number of class types and hence a large varieties of class types found in a dynamically growing dataset will lead to inaccurate classification results; and (3) An ML technique is developed based on a single learning task, and thus they are not suitable for today's multiple learning tasks and knowledge transfer requirements of Big Data analytics.

Among the ML techniques, the supervised algorithms (i.e. classification algorithms) can help to classify the network traffic data for intrusion prediction. Many supervised learning algorithms have been developed for the classification of network intrusion traffic [6-10] among them the Support Vector Machine (SVM) received a greater attention. However the computational cost of the SVM is in general higher than many other classification techniques. To ease this problem many SVM techniques have been subsequently developed in the machine learning research [11], [12]. Due to their computational complexity, they are not suitable for Big Data analytics. However the classification accuracies that the SVM techniques give are excellent. Therefore adopting SVM techniques is highly preferred. Now the challenge is to find the solution to improve the SVM technique.

4.2 Representation Learning

Representation-learning algorithms [13-15] can help supervised learning techniques to achieve high classification accuracy with computational efficiency. They transform the data, while preserving the original characteristics of the data, to another domain so that the classification algorithms can improve accuracy, reduce computational complexity and increase processing speed. However the Big Data classification requires multi-domain, representation-learning (MDRL) technique because of its large and growing data domain. The MDRL technique includes feature variable learning, feature extraction learning and distance-metric learning. Several representation-learning techniques have been proposed in machine learning research, but the recently proposed cross-domain, representation-learning (CDRL) technique by Tu and Sun [14] maybe suitable for the Big Data classification along with the suggested network model. The implementation of the CDRL technique to Big Data classification will encounter several challenges, including the difficulty in selecting relevant features, constructing geometric representation, extracting suitable features and separating the various types of data. Recently the concept of unit-circle algorithm (UCA) [15] has been proposed. It represents the intrusion traffic data by unit circles and assigns many related records to fewer unit-circles. This property can help the Big Data classification to work effectively. The public NSL-KDD dataset [16] was used in this study. The challenge now is to adopt this new algorithm to work with the modern network technology.

4.3 Machine Lifelong Learning

The continuity parameter of Big Data introduces the problems that need to be addressed by lifelong learning techniques. The learning of Big Data characteristics in short-term may not be suitable for long-term. Hence the machine lifelong learning (ML3) techniques should be used [17-19]. The concept of ML3 provides a framework that is capable of retaining learned knowledge with training examples throughout the learning phases. These features of an ML3 framework can be integrated in the technologies deployed in Figure 3. However the implementation of this framework for handling Big Data classification will have to face several challenges. One of the challenges, as stated in [19], is the scalability which is an important requirement for Big Data applications. With the suggested network topology the scalability maybe achieved but the communication challenges will add difficulties to get the data transmitted in time. Another challenge is the validation of a learned knowledge and its suitability to a new data so that the learning process is not repeated unnecessarily.

5. USER INTERACTION WITH BIG DATA

Another challenge in Big Data classification using the suggested model is the real-time access of data and processing steps. Detecting the interaction between the Big Data parameters cardinality, continuity and complexity is challenging, which requires user interaction. If the machine lifelong learning is adopted in the system then the addition of user interaction to it will help the Big Data classification significantly [4].

5.1 Data Visualization

The characteristics of Big Data make the data visualization a challenging task. As stated in [4] the recent visualization techniques like dimension reduction and data projection can only give an abstract view of the data. The abstract view, in many cases, does not give true geometric representations for the data.

This is a challenge. However, in the suggested model, the unit-circle algorithm can provide unit-circle representations to both regular and intrusion traffic, which may reduce the problem of Big Data visualization by mapping large numbers of data points to a unit-circle. If the challenge is tackled it can help the UILS unit (i.e. the user interaction and learning) to make appropriate data storage and transmission decisions.

5.2 Data Uncertainty

The communication problems during the transfer between the NTRS, HDFS and CCSS units in the network topology will lead to delay in the data or loss of data. This will lead to missing-data problem and in turn the data uncertainty issues will occur in UILS. This problem adds more complexity to the complexity parameter of the Big Data definition. Therefore, as stated in [4], the learning techniques presented earlier must consider the development of accurate knowledge from the incomplete data. Therefore the UILS unit must be built to handle this problem with user interaction and it gives significant challenges to the users.

6. CONCLUSION

This paper suggested an integration of modern technologies, Hadoop Distributed File Systems and Cloud Technologies, with the latest representation-learning technique and support vector machine to predict network intrusions through Big Data classification strategy. Additionally it suggested adopting machine lifelong learning framework for solving the problems associated with the continuity parameter. It also discussed the problems and challenges that the Big Data classification system for network intrusion prediction have to experience during the Big Data analytics. Further it suggested a change to the basic definition V^3 of Big Data to C^3 so that the Big Data analytics maybe better explained and understood with mathematical and statistical techniques. Research on Big Data techniques and technologies evolving and at the same time new problems and challenges are emerging, hence the hope is to develop better and better techniques and technologies towards finding solutions for Big Data classification problem.

7. REFERENCES

[1] P. C. Zikopoulos, C. Eaton, D. deRoos, T. Deutsch, and G. Lapis, Understanding big data – Analytics for enterprise class Hadoop and streaming data, McGraw-Hill, 2012.

[2] T. White, Hadoop: The definitive guide. O'Reilly Media, 2012.

[3] S. Carlin and K. Curran. "Cloud Computing Technologies." International Journal of Cloud Computing and Services Science (IJ-CLOSER) 1.2: 59-65, 2012.

[4] P. C. Wong, H. W. Shen, C. R. Johnson, C. Chen, and R. B. Ross, "The Top 10 Challenges in Extreme-Scale Visual Analytics," Computer Graphics and Applications, IEEE, 32(4), 63-67, 2012.

[5] I. Muttik and C. Barton. "Cloud security technologies," information security technical report 14.1: 1-6, 2009.

[6] S. B. Kotsiantis, "Supervised machine learning: A review of classification techniques," Informatica 31, 249-268, 2007.

[7] P. Laskov, C. Schafer and I. Kotenko, "Intrusion detection in unlabeled data with quarter-sphere support vector machines," In Proc. of the DIMVA Conference, 71-82, 2004.

[8] G. Huang, H. Chen, Z. Zhou, F. Yin and K. Guo, "Two-class support vector data description," Pattern Recognition, 44, 320-329, 2011.

[9] I. Corona, G. Giacinto and F. Roli, "Intrusion detection in computer systems using multiple classifier systems," Studies in Computational Intelligence (SCI) 126, 91-113, 2008.

[10] G. Giacinto, R. Perdisci and F. Roli, "Network intrusion detection by combining one-class classifier," In: F. Roli and S. Vitulano (Eds.) ICIAP 2005, LNCS 3617, 58-65, 2005.

[11] O. L. Mangasarian and D. R. Musicant, "Lagrangian support vector machine classification," TR 00-06, Data Mining Institute, Department of Computer Science, University of Wisconsin, USA – ftp://ftp.cs.wisc.edu/pub/dmi/tech-reports/00-06.pdf, 2000.

[12] V. Jeyakumar, G. Li and S. Suthaharan, "Support vector machine classifiers with uncertain knowledge sets via robust convex optimization," *Optimization* – The Journal of Mathematical Programming and Operations Research. Taylor & Francis, DOI:10.1080/02331934.2012.703667, 1-18, 2012.

[13] Y. Bengio, A. Courville and P. Vincentar, "Representation Learning: A Review and New Perspectives," arXiv:1206.5538v2 [cs.LG], 2012.

[14] W. Tu and S. Sun, "Cross-domain representation-learning framework with combination of class-separate and domain-merge objectives," In Proc. of the CDKD'12 Conference, 18-25, 2012.

[15] S. Suthaharan, "A unit-circle classification algorithm to characterize back attack and normal traffic for intrusion detection," In Proc. of the IEEE International Conference on Intelligence and Security Informatics, 150-152, 2012.

[16] NSL-KDD. http://www.iscx.ca/NSL-KDD/.

[17] S. Thrun, "Lifelong learning: A Case Study," Technical Report CMU-CS-95-208, Carnegie Mellon University, Computer Science Department, Pittsburgh, PA, 1995.

[18] D. L. Silver and R. Poirier, "Requirements for machine lifelong learning," In Proc. of the IWINAC'07, 313–319, Springer-Verlag, 2007.

[19] D. L. Silver, "Machine lifelong learning: challenges and benefits for artificial general intelligence," Lecture Notes in Computer Science, vol. 6830, 370-375, Springer, 2011.