International Journal of Information and Computation Technology.
ISSN 0974-2239 Volume 4, Number 1 (2014), pp. 33-40

http://www. irphouse.com /ijict.htm

Big Data Analytic and Mining with Machine Learning Algorithm

Jainendra Singh

Department of Computer Science, Maharaja Surajmal Institute
C-4, Janakpuri, New Delhi, INDIA.

Abstract

Big Data concern large-volume, complex, growing data sets with multiple, autonomous sources. With the fast development of networking, data storage, and the data collection capacity, Big Data are now rapidly expanding in all science and engineering domains, including physical, biological and biomedical sciences. This data-driven model involves demand-driven aggregation of information sources, mining and analysis, user interest modeling, and security and privacy considerations. We analyze the challenging issues in the data-driven model and also in the Big Data revolution.
Advances in Machine Learning (ML) provide new challenges and solutions to the security problems encountered in applications, technologies and theories. Machine Learning (ML) techniques have found widespread applications and implementations in security issues. Many ML techniques, approaches, algorithms, methods and tools are extensively used by security experts and researchers to achieve better results and to design robust systems.

Keywords: Big Data; Data Mining; Machine Learning Algorithm; Mahout.

1. Introduction

The simplest definition of "big data" is exactly what it sounds like: massive amounts (think petabytes, exabytes, zettabytes and beyond). A zettabyte is equal to over a trillion gigabytes (1,099,511,627,776 GB, to be exact or 10^{21} bytes). That's a lot of data by anybody's standards. Of course, the amount of data that constitutes "big" changes over time.

Big data typically refers to the following types of data:

- Traditional enterprise data – includes customer information from CRM systems, transactional ERP data, web store transactions, and general ledger data.
- Machine-generated /sensor data – includes Call Detail Records ("CDR"), weblogs, smart meters, manufacturing sensors, equipment logs (often referred to as digital exhaust), and trading systems data.
- Social data – includes customer feedback streams, micro-blogging sites like Twitter, and social media platforms like Facebook.

The McKinsey Global Institute estimates that data volume is growing 40% per year, and will grow 44x between 2009 and 2020. But while it's often the most visible parameter, volume of data is not the only characteristic that matters. In fact, there are four key characteristics that define big data:

- **Volume.** Machine-generated data is produced in much larger quantities than nontraditional data. For instance, a single jet engine can generate 10TB of data in 30 minutes. With more than 25,000 airline flights per day, the daily volume of just this single data source runs into the Petabytes. Smart meters and heavy industrial equipment like oil refineries and drilling rigs generate similar data volumes, compounding the problem.
- **Velocity.** Social media data streams – while not as massive as machine-generated data –produce a large influx of opinions and relationships valuable to customer relationship management. Even at 140 characters per tweet, the high velocity (or frequency) of Twitter data ensures large volumes (over 8 TB per day).
- **Variety.** Traditional data formats tend to be relatively well defined by a data schema and change slowly. In contrast, non-traditional data formats exhibit a dizzying rate of change. As new services are added, new sensors deployed, or new marketing campaigns executed, new data types are needed to capture the resultant information**.**
- **Value.** The economic value of different data varies significantly. Typically there is good information hidden amongst a larger body of non-traditional data; the challenge is identifying what is valuable and then transforming and extracting that data for analysis.

2. Analyze Big Data

Since data is not always moved during the organization phase, the analysis may also be done in a distributed environment, where some data will stay where it was originally stored and be transparently accessed from a data warehouse. The infrastructure required for analyzing big data must be able to support deeper analytics such as statistical analysis and data mining, on a wider variety of data types stored in diverse systems; scale to extreme data volumes; deliver faster response times driven by changes in behavior; and automate decisions based on analytical models. Most

importantly, the infrastructure must be able to integrate analysis on the combination of big data and traditional enterprise data. New insight comes not just from analyzing new data, but from analyzing it within the context of the old to provide new perspectives on old problems.

For example, analyzing inventory data from a smart vending machine in combination with the events calendar for the venue in which the vending machine is located, will dictate the optimal product mix and replenishment schedule for the vending machine.

3. Big Data Mining

3.1 Local Learning and Model Fusion for Multiple Information Sources

As Big Data applications are featured with autonomous sources and decentralized controls, aggregating distributed data sources to a centralized site for mining is systematically prohibitive due to the potential transmission cost and privacy concerns. On the other hand, although we can always carry out mining activities at each distributed site, the biased view of the data collected at each site often leads to biased decisions or models.

Under such a circumstance, a Big Data mining system has to enable an information exchange and fusion mechanism to ensure that all distributed sites (or information sources) can work together to achieve a global optimization goal. Model mining and correlations are the key steps to ensure that models or patterns discovered from multiple information sources can be consolidated to meet the global mining objective. More specifically, the global mining can be featured with a two-step (local mining and global correlation) process, at data, model, and at knowledge levels. At the data level, each local site can calculate the data statistics based on the local data sources and exchange the statistics between sites to achieve a global data distribution view. At the model or pattern level, each site can carry out local mining activities, with respect to the localized data, to discover local patterns. By exchanging patterns between multiple sources, new global patterns can be synthesized by aggregating patterns across all sites. At the knowledge level, model correlation analysis investigates the relevance between models generated from different data sources to determine how relevant the data sources are correlated with each other, and how to form accurate decisions based on models built from autonomous sources.

3.2 Mining from Sparse, Uncertain, and Incomplete Data

Sparse, uncertain, and incomplete data are defining features for Big Data applications. Being sparse, the number of data points is too few for drawing reliable conclusions. This is normally a complication of the data dimensionality issues, where data in a high-dimensional space (such as more than 1,000 dimensions) do not show clear trends or distributions. For most machine learning and data mining algorithms, high-dimensional spare data significantly deteriorate the reliability of the models derived from the data. Common approaches are to employ dimension reduction or feature selection to reduce

the data dimensions or to carefully include additional samples to alleviate the data scarcity, such as generic unsupervised learning methods in data mining.

Uncertain data are a special type of data reality where each data field is no longer deterministic but is subject to some random/error distributions. This is mainly linked to domain specific applications with inaccurate data readings and collections. For example, data produced from GPS equipment are inherently uncertain, mainly because the technology barrier of the device limits the precision of the data to certain levels (such as 1 meter). As a result, each recording location is represented by a mean value plus a variance to indicate expected errors. For data privacy related applications, users may intentionally inject randomness/errors into the data to remain anonymous.

This is similar to the situation that an individual may not feel comfortable to let you know his/her exact income, but will be fine to provide a rough range like [120k, 160k]. For uncertain data, the major challenge is that each data item is represented as sample distributions but not as a single value, so most existing data mining algorithms cannot be directly applied. Common solutions are to take the data distributions into consideration to estimate model parameters. For example, error aware data mining utilizes the mean and the variance values with respect to each single data item to build a Naive Bayes model for classification. Similar approaches have also been applied for decision trees or database queries. Incomplete data refer to the missing of data field values for some samples. The missing values can be caused by different realities, such as the malfunction of a sensor node, or some systematic policies to intentionally skip some values (e.g., dropping some sensor node readings to save power for transmission). While most modern data mining algorithms have in-built solutions to handle missing values (such as ignoring data fields with missing values), data imputation is an established research field that seeks to impute missing values to produce improved models (compared to the ones built from the original data).

3.3.3 Mining Complex and Dynamic Data

The rise of Big Data is driven by the rapid increasing of complex data and their changes in volumes and in nature. Documents posted on WWW servers, Internet backbones, social networks, communication networks, and transportation networks, and so on are all featured with complex data. While complex dependency structures underneath the data raise the difficulty for our learning systems, they also offer exciting opportunities that simple data representations are incapable of achieving. For example, researchers have successfully used Twitter, a well-known social networking site, to detect events such as earthquakes and major social activities, with nearly real-time speed and very high accuracy. In addition, by summarizing the queries users submitted to the search engines, which are all over the world, it is now possible to build an early warning system for detecting fast spreading flu outbreaks. Making use of complex data is a major challenge for Big Data applications, because any two parties in a complex network are potentially interested to each other with a social connection. Such a connection is quadratic with respect to the number of nodes in the network, so a million node networks may be subject to one trillion connections. For a

large social network site, like Facebook, the number of active users has already reached 1 billion, and analyzing such an enormous network is a big challenge for Big Data mining. If we take daily user actions/interactions into consideration, the scale of difficulty will be even more astonishing.

4. Big Data Challenges

There are many future important challenges in Big Data management and analytics that arise from the nature of data: large, diverse, and evolving. These are some of the challenges that researchers and practitioners will have to deal during the next years:

- Analytics Architecture: It is not clear yet how an optimal architecture of analytics systems should be to deal with historic data and with real-time data at the same time. An interesting proposal is the Lambda architecture of Nathan Marz [4]. The Lambda Architecture solves the problem of computing arbitrary functions on arbitrary data in real-time by decomposing the problem into three layers: the batch layer, the serving layer, and the speed layer. It combines in the same system Hadoop for the batch layer, and Storm for the speed layer. The properties of the system are: robust and fault tolerant, scalable, general, and extensible, allows ad hoc queries, minimal maintenance, and debuggable.
- Statistical Significance: It is important to achieve significant statistical results, and not be fooled by randomness.
- Distributed Mining: Many data mining techniques are not trivial to paralyze. To have distributed versions of some methods, a lot of research is needed with practical and theoretical analysis to provide new methods.
- Time Evolving Data: Data may be evolving over time, so it is important that the Big Data mining techniques should be able to adapt and in some cases to detect change first.
- Compression: Dealing with Big Data, the quantity of space needed to store it is very relevant. There are two main approaches: compression where we don't lose anything or sampling where we choose what is the data that is more representative. Using compression, we may take more time and less space, so we can consider it as a transformation from time to space. Using sampling, we are losing information, but the gains in space may be in orders of magnitude.
- Visualization: A main task of Big Data analysis is how to visualize the results. As the data is so big, it is very difficult to find user-friendly visualizations.
- Hidden Big Data: Large quantities of useful data are getting lost since new data is largely untagged file based and unstructured data. The 2012 IDC study on Big Data [3] explains that in 2012, 23% (643 exabytes) of the digital universe would be useful for Big Data if tagged and analyzed. However, currently only 3% of the potentially useful data is tagged, and even less is analyzed.

5. Proposed Research Work

Unfortunately traditional analytics tools are not well suited to capturing the value hidden in Big Data. The volume of data is too large for comprehensive analysis. The range of potential correlations and relationships between disparate data sources, from back end customer databases through to live web based click streams, are too great for any analyst to test all hypotheses and derive all the value buried in the data.

Machine learning is a rather new domain of IT and advanced mathematics, based on new statistical algorithms that could analyze big volume of diverse data sources (image, sound, video, social network, geo-localization, "traditional" structured database, etc…) in near real time. Computers, using these new types of programs, could learn from data for better future use.

Whether it is health, education, trade or the environment, statistical machine learning allows to analyses and gives insight in different use cases even further. In fact, machine learning algorithms are used in very diverse contexts: to recognize hand-written text, to extract information from images, to build automatic language-translation systems, to predict the behavior of customers in an online shop, to find genes that might be related to a particular disease, and so on. Generally speaking, machine learning algorithms can always be used, if we want to extract "patterns" from complex and large volume of data. In health, a lot of data are already stored on patients in various formats and it represents a huge data volume. In medical imaging, machine learning allows to see many things that we cannot see before. For example, coupling visual recognition appliance with these new ways to analyze big data helps doctors to monitor automatically elder people to see if they will fall or not (at home, in hospital, even in street).

Network security is uniquely a Big Data problem. Machines on a network generate tons of data every day—within enterprises, one terabyte of data is easily generated daily. Such a large volume practically prevents commercial security tools from performing long-range analysis, such as base-lining network object behavior over a 30 day period or more for all objects on the network. Large data volume hampers researchers' ability to perform data mining experiments to gain necessary insights.

Several approaches to machine learning are used to solve problems. The main focus will be on the two most commonly used ones —*supervised* and *unsupervised* learning—because they are the main ones supported by Mahout. Supervised learning is tasked with learning a function from labeled training data in order to predict the value of any valid input. Common examples of supervised learning include classifying e-mail messages as spam, labeling Web pages according to their genre, and recognizing handwriting. Many algorithms are used to create supervised learners, the most common being neural networks, Support Vector Machines (SVMs), and Naive Bayes classifiers. Unsupervised learning, as you might guess, is tasked with making sense of data without any examples of what is correct or incorrect. It is most commonly used for clustering similar input into logical groups. It also can be used to reduce the number of dimensions in a data set in order to focus on only the most useful attributes,

or to detect trends. Common approaches to unsupervised learning include k-Means, hierarchical clustering, and self-organizing maps.

Apache Mahout is a new open source project by the Apache Software Foundation (ASF) with the primary goal of creating scalable machine-learning algorithms that are free to use under the Apache license. The project is entering its second year, with one public release under its belt. Mahout contains implementations for clustering, categorization, and evolutionary programming. Furthermore, where prudent, it uses the Apache Hadoop library to enable Mahout to scale effectively in the cloud.

6. Conclusion

While 2012 has been the year of Big Data, 2013-14 is becoming the year of Big Data analytics. Gathering and maintaining large collections of data is one thing, but extracting useful information from these collections is even more challenging. We discussed in this paper some insights about the topic, and what we consider are the main concerns and the main challenges for the future. At the edge of statistics, computer science and emerging applications in industry, this research community focuses on the development of fast and efficient algorithms for real-time processing of data with as a main goal to deliver accurate predictions of various kinds. Machine learning techniques can solve such applications using a set of generic methods that differ from more traditional statistical techniques.

References

[1] Xindong WU, Gong-Qing WU and Wei Ding, “Data Mining with Big Data,” IEEE Transaction on Knowledge and Data Engineering, vol. 26, no. 1, pp. 97-107, Dec. 2012.
[2] Wei Fan, Albert Bifet,“Mining Big Data: Current Status and Forecast to Future,” SIGKDD Exploration, vol. 14,Issue 2,2013.
[3] J. Gantz and D. Reinsel. IDC: The Digital Universe in 2020: Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East. Dec. 2012
[4] N. Marz, J. Warren. Big Data: Principles and best practices of scalable realtime data systems. Manning Publications,2013.
[5] An Oracle White Paper June 2013
[6] Defending Networks with Incomplete Information: A Machine Learning Approach, BlackHat Briefings USA 2013
[7] A Trend Micro White Paper | September 2012
[8] Large-Scale Adaptive Machine Learning for Security Analytics, Ling Huang, Joint work with ISTC and McAfee Labs ISTC Summer Retreat, 05/31/2013
[9] http://www.skytree.net/machine-learning
[10] *hadoop.apache.org/*
[11] *mahout.apache.org/*
[12] http://www.ibm.com/developerworks/library/j-mahout/

[13] http://www.networkworld.com/community/blog/defining-big-data-security-analytics
[14] Douglas, Laney. "3D Data Management: Controlling Data Volume, Velocity and Variety". Gartner. Retrieved 6 Feb 2001.
[15] Beyer, Mark. "Gartner Says Solving 'Big Data' Challenge Involves More Than Just Managing Volumes of Data". Gartner. Archived from the original on 10 July 2011. Retrieved 13 July 2011.
[16] The White Book of BIG DATA, FUJITSU
[17] www.mckinsey.com/mgi/publication/big_data