International Journal of Engineering and Future Technology,
Year 2019; Volume 16; Issue No. 2;
ISSN 2455-6432

Challenges and Opportunities in Analytics of Big Data

Zlatinka Kovacheva

Department of Mathematics and Applied Sciences, Middle East College
P.B. No. 79, Al Rusayl, PC: 124, Muscat (Oman)
Email : zlatinka@mec.edu.om

ABSTRACT

In this paper, a statistical analysis vs. machine learning approach is presented. The main drawbacks of the classical statistical techniques are emphasized. The 3-staged process of data mining and its main methods have been discussed. The focus lays at the main advantages of neural networks in data analysis process. The conclusion is that every method for big data analysis has its application according to the specifics of the data and its required analysis.

Keywords: Big data, data mining, neural networks.

Mathematics Subject Classification: 94-06

1. INTRODUCTION

According to Asigra, a Cloud Backup company since 1986, staggering 90% of the data in the world today have been created during only two years. And, it is predicted that the worldwide number of Internet Protocol (IP) addresses will quadruple very soon. Moreover, it is forecasted three billion people will be online creating close to eight zeta bytes of data for two years only. Sixty per cent of the companies go out of business within six months after a disaster or major data loss [1].

We are living now in unprecedented era of innovative technologies that create colossal volumes of both structured and unstructured data. We need to balance between quantity and quality. We have to focus more and more at the understanding the value of Big Data and analytics. Recognizing, understanding, and using Big data in terms of scientific research are necessary at this time in a world of ever increasing data.

Big data analytics is the use of advanced analytic techniques against very large, diverse data sets that include different types such as structured/unstructured and streaming/batch, and different sizes - from terabytes to zeta bytes. Big data is a term applied to data sets whose size or type is beyond the ability of traditional relational databases to capture, manage, and process the data with low-latency. And it has one or more of the following characteristics – high volume, high velocity, or high variety, high veracity, high value, and high vitality. Big data comes from different sources - sensors, devices, video/audio, networks, log files, transactional applications, web, and social media - much of it generated in real time and in a very large scale.

Analyzing big data allows analysts, researchers, and business users to make better and faster decisions using data that was previously inaccessible or unusable. Using advanced analytics techniques such as text analytics, machine learning, predictive analytics, data mining, statistics, and natural language processing, businesses can analyze previously untapped data sources independent or together with their existing enterprise data to gain new insights resulting in significantly better and faster decisions [2].

Big data analytics examines large amounts of data to uncover hidden patterns, correlations and other insights. With today's technology, it's possible to analyze the data and get answers from it almost immediately – an effort that's slower and less efficient with more traditional business intelligence solutions [3].

Big data analytics helps organizations harness their data and use it to identify new opportunities. That, in turn, leads to smarter business moves, more efficient operations, higher profits and happier customers. In his report Big Data in Big Companies, IIA Director of Research Tom Davenport interviewed more than 50 businesses to understand how they used big data. He found they got value in the following ways [4]:

- **Cost reduction.** Big data technologies such as Hadoop and cloud-based analytics bring significant cost advantages when it comes to storing large amounts of data – plus they can identify more efficient ways of doing business;
- **Faster, better decision making.** With the speed of Hadoop and in-memory analytics, combined with the ability to analyze new sources of data, businesses are able to analyze information immediately – and make decisions based on what they've learned;
- **New products and services.** With the ability to gauge customer needs and satisfaction through analytics comes the power to give customers what they want. Davenport points out that with big data analytics, more companies are creating new products to meet customers' needs.

2. KEY TECHNOLOGIES FOR DATA ANALYTICS

We can apply the following **Key technologies** for Data Analytics [3]:

- **Data management.** Data needs to be high quality and well-governed before it can be reliably analyzed. With data constantly flowing in and out of an organization, it's important to establish repeatable processes to build and maintain standards for data quality. Once data is reliable, organizations should establish a master data management program that gets the entire enterprise on the same page.
- **Data mining.** Data mining technology helps to examine large amounts of data to discover patterns in the data – and this information can be used for further analysis to help answer complex business questions. With data mining software, we can pinpoint what's relevant, use that information to assess likely outcomes, and then accelerate the pace of making informed decisions.
- **Hadoop.** This open source software framework can store large amounts of data and run applications on clusters of commodity hardware. It has become a key technology to doing business due to the constant increase of data volumes and varieties, and its distributed computing model processes big data fast. An additional benefit is that Hadoop's open source

framework is free and uses commodity hardware to store large quantities of data. **Apache Hadoop** is based on two main concepts - Hadoop Distributed File System (HDFS) and MapReduce. Hadoop splits files into large blocks and distributes them amongst the nodes in the cluster. Hadoop MapReduce transfers packaged code for nodes to process the data [3]. Basically MapReduce refers to two separate tasks. First is the map job, which takes a set of data and converts it into another set of data, where individual elements are broken down into tuples (key/value pairs). The second one - the reduce job takes the map result and combines those data tuples into a smaller set of tuples [5].

- **Hive.** Apache Hive is an open-source data warehouse infrastructure built on top of Hadoop for providing data summarization, query, and analysis. It could also be used for analysis of data stored in HDFS system. In that case HiveQL (SQL-like language) is used to generate MapReduce code instead of writing MapReduce programs in Java [6].
- **In-memory analytics.** By analyzing data from system memory (instead of from a hard disk drive), we can derive immediate insights from the data and act on them quickly. This technology is able to remove data prep and analytical processing latencies to test new scenarios and create models; it's not only an easy way for organizations to stay agile and make better business decisions, it also enables them to run iterative and interactive analytics scenarios.
- **Predictive analytics.** Predictive analytics technology uses data, statistical algorithms and machine-learning techniques to identify the likelihood of future outcomes based on historical data. It's all about providing the best assessment on what will happen in the future, so organizations can feel more confident that they're making the best possible business decision. Some of the most common applications of predictive analytics include fraud detection, risk, operations and marketing.
- **Text mining.** With text mining technology, we can analyze text data from the WEB, comment fields, books and other text-based sources to uncover insights we hadn't noticed before. Text mining uses machine learning or natural language processing technology to comb through documents – emails, blogs, Twitter feeds, surveys, competitive intelligence and more – it helps to analyze large amounts of information and discover new topics and term relationships.

3. STATISTICAL ANALYSIS VS. MACHINE LEARNING APPROACH

For most types of experiments, sampling data is sufficient to build an effective picture of the entire dataset and, statistically, we can give high levels of accuracy to predictions based on relatively small samples. Data collected in this way is often of very high quality. To ensure the sample is representative and accurate, the data is collected and 'cleaned' with great care. This extra care is often very expensive, however, and over the last few decades we have seen the costs of running large randomized control trials spiral upwards.

Instead of researchers creating a hypothesis and collecting data from samples, machine-learning algorithms plow through large data sets searching for hypotheses.

Main drawbacks of the classical statistical techniques:

- They impose restrictions on the number of input data: one is limited to a few inputs among dozens or hundreds available, imposing a priori variable selection, with all the inherent pitfalls;
- Regressions are performed using simple dependence functions (linear, logarithmic) that are not very realistic;

- The hypothesis is made that there is only one dependence function over the whole data set, instead of many distinct niches;
- Other hypotheses imposed by their underlying theories (normal distributions, equiprobabilities, uncorrelated variables) are known to be violated, but those are necessary for their good operation;
- The necessity of using hare-brained methods to transform data.

More fundamentally, to quote Professor Peter D.M. MacDonald, of McMaster University, Ontario, Canada: **Traditional approaches to statistical inference fail with large databases**, however, because with thousands or millions of cases and hundreds or thousands of variables there will be a high level of redundancy among the variables, there will be spurious relationships, and even the weakest relationships will be highly significant by any statistical test.

4. DATA MINING

4.1 Data Mining Stages

Data Mining is an analytic process designed to explore data (usually large amounts of data) in search of consistent patterns and/or systematic relationships between variables, and then to validate the findings by applying the detected patterns to new subsets of data.

In general, the process of data mining consists of three stages:

- **Initial exploration.** This stage usually starts with data preparation which may involve cleaning data, data transformations, selecting subsets of records and - in case of data sets with large numbers of variables ("fields") - performing some preliminary feature selection operations to bring the number of variables to a manageable range (depending on the statistical methods which are being considered).
- **Model building or pattern identification with validation/verification.** This stage involves considering various models and choosing the best one based on their predictive performance (i.e., explaining the variability in question and producing stable results across samples). This may sound like a simple operation, but in fact, it sometimes involves a very elaborate process. There are a variety of techniques developed to achieve that goal - many of which are based on so-called "competitive evaluation of models", that is, applying different models to the same data set and then comparing their performance to choose the best. These techniques, which are often considered the core of predictive data mining, include: Bagging (Voting, Averaging), Boosting, Stacking (Stacked Generalizations), and Meta-Learning.
- **Deployment.** That final stage involves using the model selected as best in the previous stage and applying it to new data in order to generate predictions or estimates of the expected outcome. The concept of deployment in predictive data mining refers to the application of a model for prediction or classification to new data. After a satisfactory model or set of models has been identified (trained) for a particular application, one usually wants to deploy those models so that predictions or predicted classifications can quickly be obtained for new data.

4.2 Main DM Methods

- **Associations rules.** Association Rule Analysis is an unsupervised data mining technique that looks for frequent item sets in the corporate data. This type of data mining is very common in the retail sector and is sometimes referred to as Market Basket Analysis. By analyzing what products or services previous customers have consumed, the company

can then prompt a new customer with products they might be interested in buying. Association rule mining is a popular and well researched method for discovering interesting relations between variables in large databases. It is intended to identify strong rules discovered in databases using different measures of interestingness [7].

- **Classification analysis.** Classification Analysis is a systematic approach for obtaining important and relevant information about data and metadata. It assigns items in a collection to target categories or classes. It helps to identify to which of a set of categories a specific type of data belongs [8].
- **Clustering.** Clustering analysis is used to understand the differences and the similarities within the data. It finds clusters of data objects that are similar in some sense to one another. During the process data sets are identified where the objects are grouped based on their similarity. The goal of the method is to find high-quality clusters with low inter-cluster similarity and high intra-cluster similarity. Classification analysis and Cluster analysis are closely related as the classification can be used to cluster data [10].
- **Regression.** The goal of Regression analysis is to define the dependence between variables. It assumes a one-way causal effect from one variable to the response of another variable. Independent variables can be affected by one another but it does not mean that this dependence is both ways as is the case with correlation analysis. A regression analysis can show that one variable is dependent on another but not vice versa. Regression analysis can be used to present different levels of customer satisfactions in order to analyze how they correspond to customer loyalty.
- **Neural Networks.** They are analytic techniques modelled after the (hypothesized) processes of learning in the cognitive system and the neurological functions of the brain. Neural Networks are capable of predicting new observations (on specific variables) from other observations (on the same or other variables) after executing a process of so-called learning from existing data. As the prediction is usually an activity of the human brain, to automate this process, it is necessary to understand “How the human brain learns?”

4.3. Main advantages of neural networks

We can outline the following advantages of neural networks comparing with other data mining techniques:

- Ability to account for any functional dependence. The network discovers (learns, models) the nature of the dependence without needing to be prompted. No need to postulate a model, etc.;
- One goes straight from the data to the model without intermediary, without recoding, without binning, without simplification or questionable interpretation;
- Insensitivity to “moderate” noise or unreliability in the data.
- No conditions on the predicted variable: it can be a Yes/No output, a continuous value, one or more classes among *n*, etc.;
- Ease of handling, much less human work than traditional analytical methods;
- No need to manually detect collinearities;
- In segmentation, the net determines by itself how many clusters there are in each class;
- Speed of use: 10 microseconds when hardwired, a few milliseconds on a 1 GHz computer;
- Spatial relations (geomarketing etc.) are easily analysed and modelled;
- The final model is continuous and derivable and lends itself easily to further work;
- The neural networks model associations and not causes;
- The neural model is validated using a number of examples that were excluded from the learning set, called the “test set”. Expected and predicted values are compared.

5. CONCLUSION

The variety and complexity of the data nowadays require diversity of methods applied for data analytics. Every method has its advantages and drawbacks. The purpose of this paper is to analyse and outline them in order to support data analysts to select and apply the most suitable techniques for their specific data sets and analytical goals.

6. REFERENCES

[1] Asigra Company:

http://www.asigra.com/problems-solve/business-continuity-disaster-recovery

[2] IBM Analytics:

https://www.ibm.com/analytics/us/en/technology/hadoop/big-data-analytics/

[3] Big Data Analytics:

https://www.sas.com/en_us/insights/analytics/big-data-analytics.html

[4] Davenport, T., Dyché J., 2013*, Big Data in Big Companies,* SAShttps://www.sas.com/en_us/whitepapers/bigdata-bigcompanies-106461.html

[5] What is MapReduce.
https://www-01.ibm.com/software/data/infosphere/hadoop/mapreduce/

[6] Oracle® Big Data Appliance,
https://docs.oracle.com/cd/E37231_01/doc.20/e36963.pdf

[7] Agrawal, Imielinski T., Swami A., 1993, *Mining Association Rules Between Sets of items in large databases,* Proceedings of the ACM SIGMOD Conference on Management of data, 207-216.

[8] Peralta D. at all, 2015, *Evolutionary Feature Selection for Big Data Classification: A MapReduce Approach,* Mathematical Problems in Engineering, Vol. 2015, Hindawi Publishing Corporation

[9] Han J., Kamber M., 2006, *Data Mining: Concepts and Techniques*, Morgan Kaufmann, San Francisco, 2nd edition.

[10] Min Ji, Xie F., Ping Y., 2013, *A Dynamic Fuzzy Cluster Algorithm for Time Series, Abstract and Applied Analysis*, vol. 2013, Article ID 183410

[11] Laney D., 2012, *The Importance of 'Big Data': A Definition*, Gartner,