Seminar Predictive Analytics in Big Data WS 2016/17

Jasim Waheed Ansari

RWTH Aachen University, 52056 Aachen, Germany
jasim.waheed@rwth-aachen.de

Abstract. Predictive Analytics (PA) has replaced the idea of ad-hoc data analysis by fact based decisioning. PA include statistical methods from machine learning and data mining to build data model for regression, prediction, neural networks, classification purposes. In order to incorporate big data criterion, the frameworks and tools that perform modeling using previously stated methods has to address the 4Vs of Big Data, namely volume, veracity, velocity and value. This paper is aimed at providing an overview of big data analytics framework and then highlighting few major tools, comparison and assessment of their characteristics. The second aim is to go through the major issues and challenges that are faced while carrying out PA in big data. We will also address the possible solutions in form of best practices concerning those problems. The third aim is to provide an example from ERP systems to highlight its predictive analytics capabilities using big data systems. The fourth aim is to present a conceptual framework of integrating Complex Event Processing and PA called Predictive Complex Event Processing. We will observe that the given framework would be a generic design pattern for future work. On an ending note, we will demonstrate a case study to get a lucid idea on a practical level.

1 Introduction

In recent years, there is a flood of data that is shared or generated from various sources. This data is usually presented in an unstructured format such as videos, audios, texts, images. The presented data could or could not be related to each other or there could be a possibility of having some hidden patterns. To perform any sort of analysis, analysts cannot use such unstructured data directly. Such data requires conversion into a well-formed format (similar to a relational data), which could be used by data scientists for purpose of analysis on them. As a result, this structured data can then be used to decipher hidden meanings or pattern which supports prediction of market behavior, enterprise needs, enabling precision based decisioning using technique called Predictive Analytics.

1.1 What is Predictive Analytics?

Predictive Analytics is the process of discovering artifacts or useful information from a set of data (structured,unstructured or semi-structured) in order to make

predictions and assessment about future results. It surrounds techniques from statistics, data mining, machine learning and artificial intelligence.

Predictive Analytics steps: As per general guideline, the steps along with percentage of time spent on each of the step are as follows:

1. Understanding the domain (5-10 percent)
2. Understanding the data (5-10 percent)
3. Preparing the data (50-60 percent)
4. Modeling (5-15 percent)
5. Evaluation (5-10 percent)
6. Deployment (10-15 percent)

It is worth noting that each step could have as many iterations as needed. Adding to that, core effort in processing is emphasized at the data.
Figure 1 shows Predictive Analytics process which is followed actively by analysts across various industries [6]

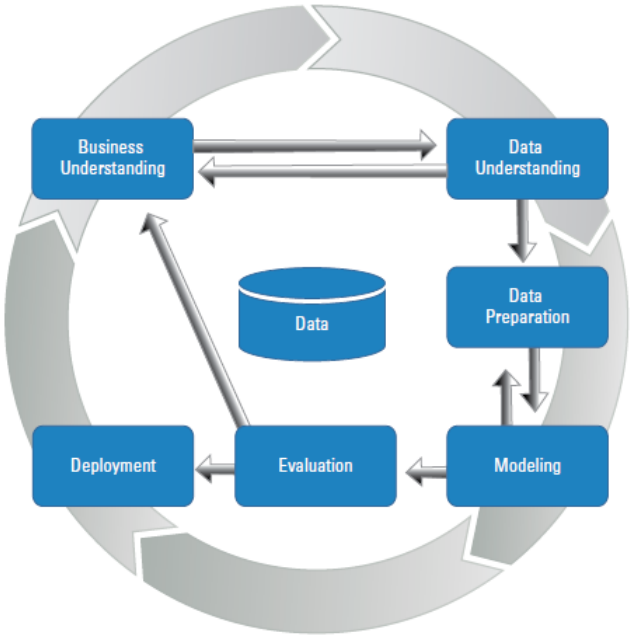

Fig. 1. Predictive analytics process

Techniques available: From predictive analytics process, modeling and evaluation steps are where analysts use algorithms to produce key information they desire. Predictive modeling algorithms [7] are divided into two groups: *Supervised Learning* and *Unsupervised Learning*.
In Supervised Learning algorithms, the predictor value or target variable is *supervisor*, it is the column in dataset that is used for predicting value from other column values. The principal algorithms under supervised learning are *regression* for continuous target variable and *classification* for class based on target variable. Supervised learning is also called sometimes as *predictive modeling*.
In Unsupervised Learning algorithms, which is also sometimes called as *descriptive modeling*. It contains no target variable or class label. Analysis is done by grouping or clustering the dataset based on proximities of other input values. Each cluster is given a label name to uniquely identify from other clusters.

1.2 Predictive Analytics: Harness the power of Big Data

Almost every decision making process, be it in any enterprise, involves predictive analytics to drive their business better and helps gaining competitive edge in the market. Decision making in current era is based on day-to-day operational business data rather than on only special projects or scenario. Current tools and technologies are unsatisfactory and not up to standards to process such huge chunks of operational data. They are also unable to make in-depth insight and generate value.
Thanks to Big Data technology and tools, predictive analytics can be applied to deluge of data at enterprise level. We have now many ways to tackle and test different predictive models at various level of framework for business strategies. Contrarily, there are also side-effects if the tools are not applied correctly, it could lead to loss of business value and shares at incredibly surprising pace.
In the upcoming sections, we will be providing the conjunction of Predictive Analytics with Big Data technology and bring the following topics in coherence:
(1) understanding big data architecture for analytical perspective
(2) comparison and assessment of big data analytical tools used for predictions
(3) address few of the challenges which collides interest of predictive analytics with big data. One of the key challenges being privacy- arising from ever increasing data from online services and personal devices such as mobile phones. These are subjected to risk of personal information being exposed for illicit uses.
(4) discuss example from Enterprise Resource Planning (ERP) systems, and explore the opportunities of predictive analytics on ERP system's integrated big data hub.
(5) discuss about Complex Event Processing which deals in identifying complex events based on the rules dictated by users of the system. In order to avoid the manual intervention of users for providing progressive rules, we will discuss about the framework that includes Predictive Analytics technologies in Complex Event Processing tools and applications.
(6) case study demonstrating the power of predictive analytics coupled with big data technology.

References

1. Chandarana, P., Vijayalakshmi, M.: Big data analytics frameworks. In: Circuits, Systems, Communication and Information Technology Applications (CSCITA), 2014 International Conference on. (2014) 430–434
2. Babu, M.S.P., Sastry, S.H.: Big data and predictive analytics in erp systems for automating decision making process. In: Software Engineering and Service Science (ICSESS), 2014 5th IEEE International Conference on. (2014) 259–262
3. Earley, S.: Big data and predictive analytics: What's new? IT Professional **16**(1) (2014) 13–15
4. Fülöp, L.J., Beszédes, A., Tóth, G., Demeter, H., Vidács, L., Farkas, L.: Predictive complex event processing: A conceptual framework for combining complex event processing and predictive analytics. In: Proceedings of the Fifth Balkan Conference in Informatics. BCI '12, New York, NY, USA, ACM (2012) 26–31

5. Hu, H., Wen, Y., Chua, T.S., Li, X.: Toward scalable systems for big data analytics: A technology tutorial. IEEE Access **2** (2014) 652–687
6. Wessler, M.: Predictive analytics for dummies. Alteryx Special Edition. Wiley (2014)
7. Zekić-Sušac, M., Has, A.: (Predictive analytics in big data platforms–comparison and strategies)