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 predicting useful information from a set of data (structured,unstructured or semi-structured). 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]

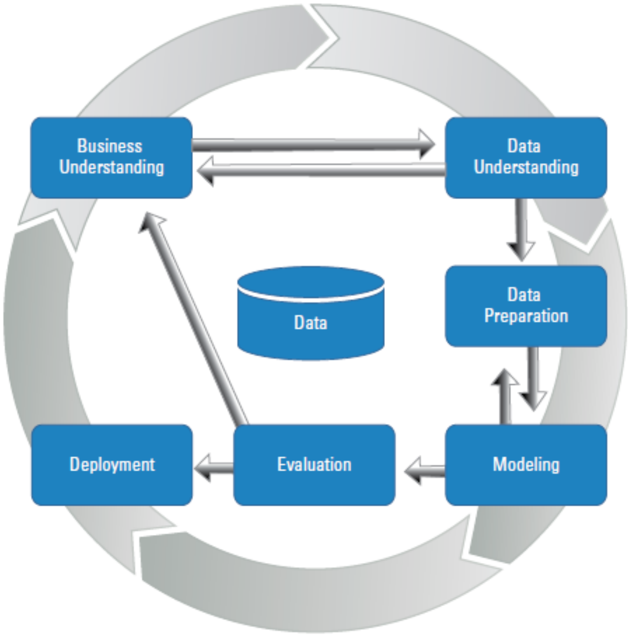

Fig. 1. Predictive analytics process

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] falls under Supervised Learning. In Supervised Learning, the predictor value or target variable is *supervisor*, it is the column in dataset that is used for predicting value from other column values. It is divided into two sets:

1. Classification: class based or discrete target variable
2. Regression: continuous target variable

Figure 2 shows Predictive Analytics techniques used for modeling[8].

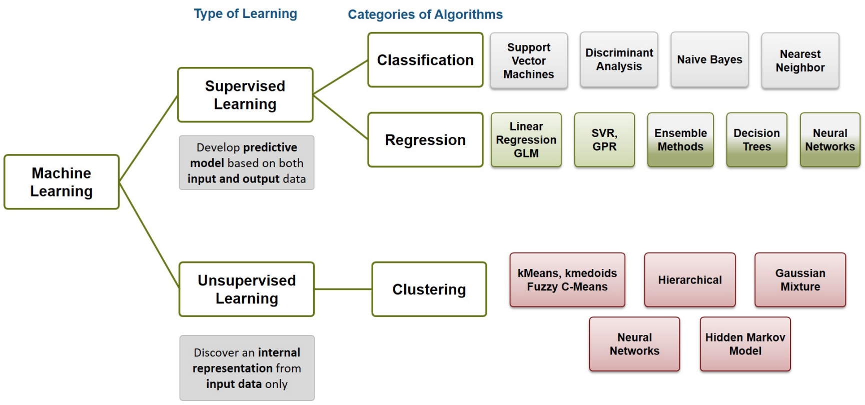

Fig. 2. Predictive analytics techniques

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, sometimes also called as Big Data Analytics. We have now many ways to tackle and test different predictive models at various level of framework for business strategies. Figure 3 shows delineation of characteristics of Predictive Analytics techniques in transactional databases vs big data technologies. The former is based on prediction through historical structured data whereas big data platform can be used for modeling in real-time with structured, unstructured or semi-structured data. The common methods involved in transactional databases are data mining algorithms such as decision trees, regression, clustering, association rules, etc. In big data mediums, advanced techniques such as speech recognition, mobile based analytics, etc. [[7]].

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.

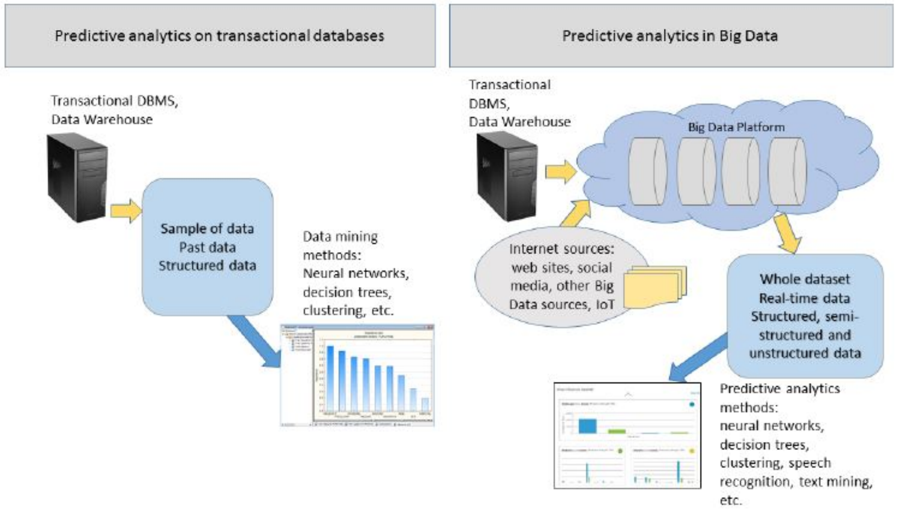
Predictive analytics on transactional databases
Predictive analytics in Big Data
Transactional DBMS, Data Warehouse
Sample of data
Past data
Structured data
Data mining methods:
Neural networks,
decision trees,
clustering, etc.
Transactional DBMS, Data Warehouse
Big Data Platform
Internet sources: web sites, social media, other Big Data sources, IoT
Whole dataset
Real-time data
Structured, semi-structured and unstructured data
Predictive analytics methods:
neural networks,
decision trees,
clustering, speech recognition, text mining, etc.

Fig. 3. Predictive analytics in traditional databases vs Big Data

2 Big Data System Architecture

A big data system is immensely complex. In order to demonstrate the architecture of predictive analytics in big data, we should firstly understand the categorization of two paradigms of big data analytics with respect to processing time requirements:

- Stream Processing: Streaming processing data is dependent on fresh data i.e. it needs to be analyzed as soon as it is available. Some famous open source systems available are Storm and Kafka.
- Batch Processing: In batch processing, data is stored as chunk and then analyzed. One of the most famous models available is Map Reduce.

In this section, we will discuss architecture based on batch processing types, since it is widely adopted in industries. We will show one such framework placed on value chain for big data analytics[5]. Big data value chain framework involves four stages i.e. generation, acquisition, storage and analytics. In this work, we will see leading technologies in analysis phase.

2.1 Big Data Value Chain architecture

This architecture is based on system engineering approach, well recognized in industries, the notion is to break down the big-data system into four continuous phases in horizontal axis as shown in Figure 4 [5].

Data Analysis: It is the main and concluding step from big data value chain that focuses on the aspect of analytics. This phase supports mechanisms or tools to investigate, transform, and modeling data to discover insights. We will discuss hereby some classification metric of big data analytics. Furthermore, we will describe common methods in big data analytics that together build the pillar for making predictions comprehensively.

Categories: Blackett et al [9] proposed classification of big data analytics into three levels as follows:

1. Descriptive Analytics: analyses historical dataset to describe what happened. It is widely used with business intelligence or ERP solutions.
2. Predictive Analytics: focuses on future predictions and trends, uses supervised learning algorithms to understand trends, and data mining exacts insight.
3. Prescriptive Analytics: focuses on decision making. For instance, simulation is used to analyze the systems and recognize various optimization techniques to get efficient solutions.

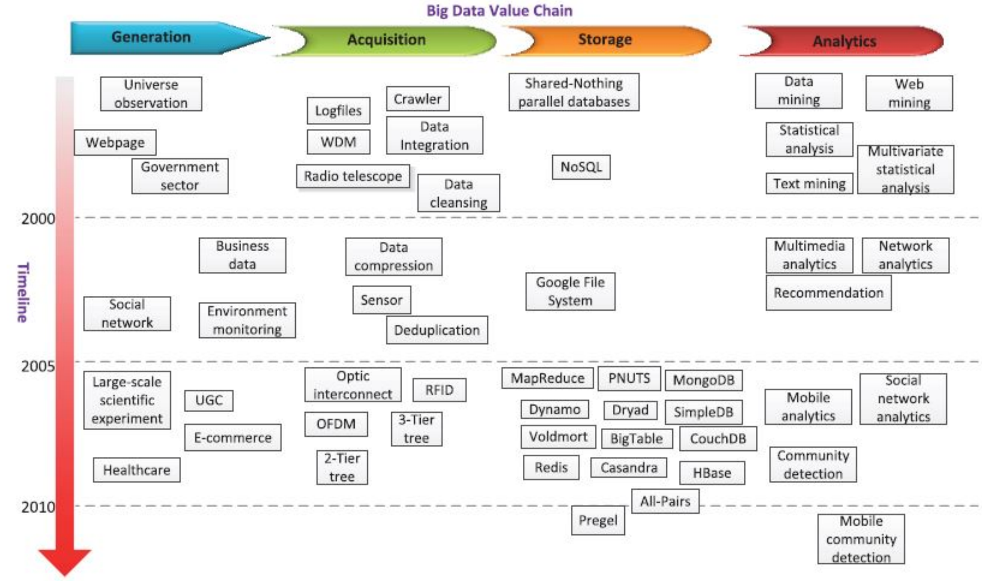

Fig. 4. Big data value chain architecture. Horizontal axis composes of four continuous phases for data life cycle. For each stage, we demonstrate cutting edge technologies available past ten years in vertical axis

Common Methods: Even though the motive, application domains varies for different types of analysis, there exists three common methods that are useful to discuss.

1. Data Visualization: It helps in information visualization using visual interface. Arming experts with such tools would help in making predictions much more smooth as it would provide easy, user-friendly access to results.

2. Statistical Analysis: It is related to statistical theory, sub-area of applied mathematics. Randomness is modeling using probability theory. This area performs inferential analysis on large data set for prediction processes.

3. Data Mining: It is the computational process of discovering patterns in large data sets.[5] Different applications applies principles and techniques from data mining such as regression, clustering, classification and link detection. Furthermore, advance algorithms such as deep learning are current trending in big data analytics.

3 Predictive Analytics Technologies

3.1 Big Data Platforms: Hadoop and Spark

3.2 Big Data Predictive Analytics Tools

3.3 Results: Comparison and Strategies

4 Issues in predictive analytics

4.1 Privacy challenges using Big Data analytics

4.2 Best practices available for preserving privacy

5 Example: Big Data Predictive Analytics in ERP Systems

5.1 ERP System Landscape

5.2 ERP with Big Data Using Predictive Analytics

5.3 Predictive Modeling tasks in ERP systems

6 Conceptual Framework: Predictive Complex Event Processing

6.1 Background

6.2 Exploiting the combined value of Complex Event Processing and Predictive Analytics

6.3 Conceptual CEP-PA framework

6.4 Proof of Concept

7 Case Study: Predicting Election Trends using Twitter: Hillary Clinton vs Donald Trump

7.1 Building the network of Twitter users

7.2 Inferring the opinion of users

7.3 Predicting Election Trends

7.4 Results

8 Conclusion

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)
8. Pilotte, P.: (Analytics-driven embedded systems, part 2 developing analytics and prescriptive controls)
9. Blackett, G.: (Analytics network-o.r. analytics)