Chapter 12
Big Data and Web Intelligence:
Improving the Efficiency on Decision Making Process via BDD

Alberto Pliego
Escuela Técnica Superior de Ingenieros Industriales, Spain

Fausto Pedro García Márquez
Escuela Técnica Superior de Ingenieros Industriales, Spain

ABSTRACT

The growing amount of available data generates complex problems when they need to be treated. Usually these data come from different sources and inform about different issues, however, in many occasions these data can be interrelated in order to gather strategic information that is useful for Decision Making processes in multitude of business. For a qualitatively and quantitatively analysis of a complex Decision Making process is critical to employ a correct method due to the large number of operations required. With this purpose, this chapter presents an approach employing Binary Decision Diagram applied to the Logical Decision Tree. It allows addressing a Main Problem by establishing different causes, called Basic Causes and their interrelations. The cases that have a large number of Basic Causes generate important computational costs because it is a NP-hard type problem. Moreover, this chapter presents a new approach in order to analyze big Logical Decision Trees. However, the size of the Logical Decision Trees is not the unique factor that affects to the computational cost but the procedure of resolution can widely vary this cost (ordination of Basic Causes, number of AND/OR gates, etc.) A new approach to reduce the complexity of the problem is hereby presented. It makes use of data derived from simpler problems that requires less computational costs for obtaining a good solution. An exact solution is not provided by this method but the approximations achieved have a low deviation from the exact.

INTRODUCTION

The information and communication technologies (ICT) have grown up with no precedents, and all aspects of human life have been transformed under this new scenario. All industrial sectors have rapidly incorporated the new technologies, and some of them have become de facto standards like supervisory control and data acquisition (SCADA) systems. Huge large amounts of data started to be created, pro-

cessed and saved, allowing an automatic control of complex industrial systems. In spite of this progress, there are some challenges not well addressed yet. Some of them are: the analysis of tons of data, as well as continuous data streams; the integration of data in different formats coming from different sources; making sense of data to support decision making; and getting results in short periods of time. These all are characteristics of a problem that should be addressed through a big data approach.

Even though Big Data has become one of the most popular buzzword, the industry has evolved towards a definition around this term on the base of three dimensions: volume, variety and velocity (Zikopoulos and Eaton, 2011).

Data *volume* is normally measured by the quantity of raw transactions, events or amount of history that creates the data volume. Typically, data analysis algorithms have used smaller data sets called training sets to create predictive models. Most of the times, the business use predictive insight that are severely gross since the data volume has purposely been reduced according to storage and computational processing constraints. By removing the data volume constraint and using larger data sets, it is possible to discover subtle patterns that can lead to targeted actionable decisions, or they can enable further analysis that increase the accuracy of the predictive models.

Data *variety* came into existence over the past couple of decades, when data has increasingly become unstructured as the sources of data have proliferated beyond operational applications. In industrial applications, such variety emerged from the proliferation of multiple types of sensors, which enable the tracking of multiple variables in almost every domain in the world. Most technical factors include sampling rate of data and their relative range of values.

Data *velocity* is about the speed at which data is created, accumulated, ingested, and processed. An increasing number of applications are required to process information in real-time or with near real-time responses. This may imply that data is processed on the fly, as it is ingested, to make real-time decisions, or schedule the appropriate tasks.

However as other authors point out, Big Data could be also classified according to other dimensions such as veracity, validity and volatility.

Data *veracity* is about the certainty of data meaning. This feature express whether data reflect properly the reality or not. It depends on the way in which data are collected. It is strongly linked to the credibility of sources. For example the veracity of the data collected from sensors depends on the calibration of sensors. The data collected from surveys could be truthful if survey samples are large enough to provide a sufficient basis for analysis. In resume, the massive amounts of data collected for Big Data purposes can lead to statistical errors and misinterpretation of the collected information. Purity of the information is critical for value (Ohlhorst, 1964).

Data *validity* is about the accuracy of data. The validity of Big Data sources must be accurate if results are wanted to be used for decision making or any other reasonable purpose (Hurwitz et al, 2013)

Data *volatility* is about how long the data need to be storage. Some difficulties could appear due to the storage capacity. If storage is limited, what and how long data is needed to be kept. With some Big Data sources, it could be necessary to gather the data for a quick analysis (Hurwitz et al, 2013).

These data are often used for decision making. DM processes are done continuously by any firm in order to maximize the profits reliability, etc. or minimize costs, risks, etc. There are software to facilitate this task, but the main problem is the capability for providing a quantitative solution when the case study has a large number of BCs. The DM problem is considered as a cyclic process in which the decision maker can evaluate the consequences of a previous decision. Figure 1 shows the normal process to solve a DM problem.

Related Content

Dynamic View Selection for OLAP

Michael Lawrenceand Andrew Rau-Chaplin (2008). *International Journal of Data Warehousing and Mining (pp. 47-61).*

www.irma-international.org/article/dynamic-view-selection-olap/1799

Semantic Integrating for Intelligent Cloud Data Mining Platform and Cloud Based Business Intelligence for Optimization of Mobile Social Networks

Gebeyehu Belay Gebremeskel, Zhongshi Heand Xuan Jing (2013). *Data Mining in Dynamic Social Networks and Fuzzy Systems (pp. 173-211).*

www.irma-international.org/chapter/semantic-integrating-intelligent-cloud-data/77528

Mining Progressive Confident Rules in Sequence Databases

Wynne Hsu, Mong Li Leeand Junmei Wang (2008). *Temporal and Spatio-Temporal Data Mining (pp. 87-111).*

www.irma-international.org/chapter/mining-progressive-confident-rules-sequence/30263

An OLAM Operator for Multi-Dimensional Shrink

Stefano Rizzi, Matteo Golfarelliand Simone Graziani (2015). *International Journal of Data Warehousing and Mining (pp. 68-97).*

www.irma-international.org/article/an-olam-operator-for-multi-dimensional-shrink/129525

Information Privacy and Security for e-CRM

Jerry Fjermestadand Nicholas C. Romano (2009). *Social Implications of Data Mining and Information Privacy: Interdisciplinary Frameworks and Solutions (pp. 48-70).*

www.irma-international.org/chapter/information-privacy-security-crm/29144