Journal of Advances and Scholarly Researches in Allied Education
Vol. XV, Issue No. 4, June-2018, ISSN 2230-7540

Data Mining for Business Decisions: A Literature Review

Parag Chandra Dutta*

Abstract – The presence of "big data", or this massive amount of increasing data, offers both an opportunity as well as a challenge to researchers. A lot of progress has been made in developing the capability to process, store, and analyze big data: In addition to the big data computing capability (in terms of processing and storing big data in a distributed fashion on a cluster of computers), the rapid advances in using intelligent data analytics techniques—drawn from the emerging areas of artificial intelligence (AI) and machine learning (ML)—provide the ability to process massive amounts of diverse unstructured data that is now being generated daily to extract valuable actionable knowledge. This provides a great opportunity to researchers to use this data for developing useful knowledge and insights.

Keywords: Big Data, Mining, Analytics

----------X----------

INTRODUCTION

Modern datasets, or the big data, differ from traditional datasets in 3 V's: volume, velocity and variety. In today's age huge volumes of data is being generated at huge pace (or velocity) and the numerous sources of data give vast variety to it. All of this data, if harnessed intelligently, can truly realize the notion of the information age.

Actionable information can be gathered from the data after performing intelligent processing and analytics on the available data. The techniques (specially related to machine learning) in order to gather, store, process and analyze this vast amount of data are the subject matter of this section. We also try to link this discussion, and different examples considered here to explain various concepts, to the humanitarian development.

A. Machine Learning

Machine learning (ML), a sub-field of artificial intelligence (AI), focuses on the task of enabling computational systems to learn from data about how to perform a desired task automatically. Machine learning has many applications including decision making, forecasting or predicting and it is a key enabling technology in the deployment of data mining and big data techniques in the diverse fields of healthcare, science, engineering, business and finance.

Broadly speaking, ML tasks can be categorized into the following major types:

1) Supervised Learning: In this class of ML, the learning task is to generalize from a training set, which is labeled by a "supervisor" to contain information about the class of an example, so that predictions can be made about new, yet unseen, examples. If the output (or prediction) belongs to a continuous set of values then such a problem is called regression, while if the output assumes discrete values then the problem is called classification. In the following we briefly present a few classification techniques.

Naive Bayes Classifiers are based on Baye's Theorem that assume independence among features given a class. These has been widely used for the Internet traffic classification: e.g., naive Bayesian classification of the Internet traffic.

Decision Trees (DT) define a popularly used intuitive method that can be used for learning and predicting about target features both for quantitative target attributes as well as nominal target attributes. Although, DT do not always perform very competitively, their main advantage is their intuitive interpretation which is crucial even network operators have to analyze and interpret the classification method and results.

Support Vector Machines (SVM) is a widely used supervised learning technique that is remarkable for being practical and theoretically sound, simultaneously. The approach of SVM is rooted in the field of statistical learning theory, and is

systematic: e.g., training a SVM has a unique solution (since it involves optimization of a concave function).

2) Unsupervised Learning Techniques: The basic method in unsupervised learning is clustering. In clustering, the learning task is to categorize, without requiring a labeled training set, examples into 'clusters' on the basis of perceived similarity. This clustering is used to find the groups of inputs which have similarity in their characteristics. Intuitively, clustering is akin to unsupervised classification: while classification in supervised learning assumed the availability of a correctly labeled training set, the unsupervised task of clustering seeks to identify the structure of input data directly.

3) Reinforcement Learning: This is a reward/ punishment based ML technique. In this technique a learner, based on an input received, performs some action, potentially affecting the environment around it. This action is then rewarded or punished. The nature of the mapping from the actions taken by the learner to rewards/ punishments, in general, is probabilistic in nature. The eventual goal of a learner is to discover such an optimal mapping (or policy), from its actions to the rewards/ punishments, so that the average long-term reward is maximized.

4) Deep Learning: Deep learning (DL) is an ML technique that comprises deep and complex architectures. These architectures consist of multiple processing layers, each capable of generating non-linear response corresponding to the data input. These layers consist of various small processers running in parallel to process the data provided. These processors are called neurons. DL has proved to be efficient in pattern recognition, image and natural language processing. DL finds its applications in very broad spectrum of applications ranging from healthcare to the fashion industry, with many key technology giants like Google, IBM and Facebook deploying DL techniques to create intelligent products.

5) Association Rule Learning: It is a method for discovering interesting relations between variables in large databases. In this, we seek to learn about associations between the features present in examples. Unlike classification (supervised learning), which strictly and discretely tells the class of an example, relations or associations among various variables in an example database are considered in association rule learning.

6) Numeric Prediction: In numeric prediction, we are not interested in predicting the discrete class (or category) to which the example belongs, but the numeric quantity associated with it. As an example consider, once again, the weather dataset mentioned to explain the association learning. Now consider the classification problem where instead of predicting whether (based on the given features) a game will be played or not a numeric quantity, e.g., how long (in minutes) a game is likely to be played, is predicted as an output.

DATA MINING FOR BUSINESS DECISIONS

Data mining usually refers to automated pattern discovery and prediction from large volumes of data using ML techniques. Data mining can also be used to refer to online analytical processing (OLAP) or SQL queries that entails retrospectively searching a large database for a specific query. OLAP queries, also known as decision-support queries, are typically complex expensive queries that take a long time and touch large amounts of data. The process of extracting useful information or knowledge from the structured/ unstructured data and databases (relational and non-relational), using data mining and ML techniques, is called knowledge discovery, sometimes collectively called KDD (knowledge discovery in databases). This knowledge can be in the form of brief and concise visual reports, a predicted value or a model of a larger data generating system.

1) New Trend in Database Technology: NoSQL: With the advent of big data and Web 2.0 we now have a huge amount of unstructured data such as word documents, email, blog posts, social- and multimedia data. This unstructured data is different from the structured data in that it cannot be stored in an organized fashion in the conventional relational databases. In order to store and access unstructured data, a different approach and techniques are required. NoSQL (or non-relational) databases have been developed for the same purpose.

Companies like Amazon and Google adopt this approach for storing and accessing their data. The main advantage, besides storing unstructured data, is that these NoSQL databases are distributed and hence easily scalable, fast and flexible (as compared to their relational counterpart). One of the concerns in using NoSQL datases, though, is that they usually do not inherently support the ACID (atomicity, consistency, integrity and durability) set, as supported by the relational databases. One has to manually program these functionality into one's NoSQL database.

2) **Predictive Analytics:** Predictive analytics refers to a technology that aims to provide a competitive advantage by predicting some future occurrences or behavior (using data mining and ML techniques) based on past experience (in the form of collected data). Predictive analytics encompasses data science, machine learning, predictive and statistical modeling and outputs empirical predictions based on given input empirical data.

Crowdsourcing and Big Data

Crowdsourcing is different from outsourcing. In crowdsourcing, the nuance is, a task or a job is outsourced but not to a designated professional or organization but to general public in the form of an open call. Crowdsourcing is a technique that can be deployed to gather data from various sources such as text messages, social media updates, blogs, etc. This data can then be harmonized and analyzed in mapping disaster struck regions and to further enable the commencement of search operations. This technique helped during the 2010 Haiti earthquake.

Internet of Things

Internet of things (IoT) is a new trendy field fueled by the hype in big data, emergence of network science, proliferation of digital communication devices and ubiquitous Internet access to common population. A technical report by McKinsey Global Institute], presents the potential of IoT in terms of economic value. According to the study , if all the challenges are overcome, the IoT has a potential to create 3$–11$ trillion USD worth of economic value. In IoT, different sensors and actuators are connected via a network to various computing systems providing data for actionable knowledge.

In this way IoT, big data and network science are all related. Interoperability, harmony of data from one system with another, is a potential challenge in the way of IoT expansion. IoT finds its application in healthcare monitoring systems. Data from wearable body sensory devices and hospital health care databases, if made interoperable, could help doctors to make more efficient decisions in diagnosing and monitoring chronic diseases. Similarly, with the help of ML techniques and predictive analytics, data that is fed in real-time to computing systems by sensors and actuators can be utilized to revolutionize the maintenance tasks in industries with a significant reduction in the breakdowns of parts and system downtimes.

BIG DATA FOR DEVELOPMENT: DEVELOPMENT AREAS

1) Natural Disasters: Meier talks about the important and crucial role that the analysis of big data can play when a natural disaster strikes a part of the world. When an earthquake hit Haiti in 2010, after this incident the community of online users played a very significant role to fight this disaster. Through crowd sourcing, a real-time image of the situation, or a crisis map, became clear. Big data techniques from the fields of AI and ML were deployed to find meaning in massive and fast-changing online data comprising of tweets and short message service (SMS), which was generated after the disaster. The author calls members of this community the digital humanitarians.

2) Migrant Crisis: As we write this paper, the ongoing political unrest in Syria, which started in 2011, worsens day after day. This situation has displaced a large number of people internally and a staggering number outside the country. The fleeing of people from the troubled areas, leaving their own homes and finding shelter elsewhere, has resulted in mass movement of population—the magnitude of which has not been observed since the end of World War II. Syria's immediate neighbors—in particular, Lebanon and Jordan—and many countries in Europe are seeing a huge influx of people in search of shelter, better and less troubled lives. In this type of scenario, it is a challenge for the humanitarian organizations to operate efficiently especially in the troubled and war-torn areas. Two major challenges are faced by the helping organizations: (i) ensuring that the right regions get the right type of assistance in time; and (ii) ensuring the coordination within and among organizations during such times. This is important to avoid chaos and mismanagement. In both of these cases, data has a vital role to play

Healthcare

Big data analytics in healthcare is bringing a huge cultural change in the way conventional medical diagnosis and treatment operates. Big data can revolutionize medical diagnosis by integrating data gathered from various medical records of a patient, as well as real-time wearable sensors, to analyze and diagnose the patient's current health status and provide an early warning sign if the health of a patient is on a dangerous track. Doing this helps in taking preventing measurements to diagnose and treat a potentially harmful disease during early stages. In terms of making treatment more efficient and convenient, it is possible for a person having a smart phone to access medical service providers via a healthcare app to obtain quick and more personalized response from the convenience of one's home.

Education

The field of education is making a transition to digital era with the use of physical textbooks waning and digital versions of study material gaining more popularity. Education is one of the fields that has greatly benefited from the big data analytics. The conventional pedagogical practices, students' learning and study habits, and the way whole educational system is being designed and run are seeing revolutionary changes. In particular, the practice on online learning and blended learning is gaining popularity. In blended learning, online teaching, learning and assessments are combined with the conventional pedagogical approach.

There are two important interrelated big data related developments in education: learning analytics and educational data mining. Learning analytics (LA) is an emerging cross-disciplinary field that combines data analytics and learning, thus bringing researchers together from various fields such as computer science, data science and social science. In this field, research is carried out for various purposes that include, but are not limited to, predictive analysis, social network and sentiment analysis, personalized learning, and better curriculum designs. Educational data mining (EDM), like LA, is also an emerging and related field. In EDM, data mining and ML techniques are applied to the data representing the student's interaction with the digital and online educational system—which can be easily stored in massive open online courses (MOOCs) and online tests—to help the students better learn.

BIG DATA ANALYTICS FOR DEVELOPMENT

Mobile Analytics

Mobile analytics is the application of big data techniques to the massive amounts of data that mobile companies gather about their users in terms of call volume, calling pattern, and location. This data contains a wealth of information that can be very useful for research, planning and development (the use of such information also poses many privacy and ethical use challenges). The field of mobile big data analytics focuses on analyzing cell-phone data to provide insights that can be used to drive value-added services. For example "call-detail-records" (CDR) analysis maintained by mobile service providers can 8 be used for gathering socioeconomic information.

Technical (strict and secure data storage and data anonymization techniques) and agreement based approaches (between volunteers and researchers) were adopted to ensure the privacy of the volunteers during both the data collection and making it available to the research community. If privacy and other issues related to big data are taken care of then these projects are very important in enabling the research efforts to explore the immense potential that the big data has to impact the future of technology.

Living Analytics

Living analytics is a big-data-driven interdisciplinary field of research that incorporates expertise from a number of disciplines including computer science, network science, social science, and statistics. Living analytics is related to the study of social and behavioral patterns of individuals and societal groups. Like other fields, social science has also advanced through the recent development in big data technology: the field of computational social science is inherently based around using the advances in storage and computing capabilities to process readily available big data for advancing our understanding of social science. Conventional social science techniques, which are mainly based on questionnaires and surveys, suffer from bias, incompleteness or sometimes inaccurate and scarce information. Modern techniques where data from devices, specially cell-phones and other digital communication devices, are collected and different models are formed to study the structure and dynamics of a social network either on individual or collective levels are in contrast with the conventional methods. Intelligent techniques are being devised and deployed to mine useful data from the massive datasets gleaned from cell-phones and other digital devices.

Visual Analytics

Visual analytics is an interesting branch of big data exploration in which the aim is to support the science of analytical reasoning through interactive visual interfaces. Through information visualization, large amounts of quantitative data can be shown in a limited space. In visual analytics there might not be much a priori information known about the data or even about the data exploration goals. In information (or data) exploration the goals are steered and fine-tuned during the process of exploration by human interaction. Visual analytics has the power to quickly convey the essence of a massive dataset to a user as contrast to automatic data mining and machine learning tools, which require more technological soundness and knowledge.

As an example, as it quite often happens these days, the viral trends on any one of the social media sites, e.g. Twitter or Facebook, can provide one with a good idea of the trend if this outbreak is shown in an animated time lapsed video. One can track the origin and hubs responsible for the spread of the virus. Through these principles, combined with the concepts from the network science, the outbreak of biological viruses can also be analyzed or even prevented beforehand.

Data maps:

This type of visualization is usually a cross of cartography and statistical information. In data maps a region of interest is considered and a specific variable under-consideration is analyzed over the spatial dimensions of this region. Quoting an example from the book, a map of the USA is considered and death-rates due to different types of cancers are shown all over the map. This statistical information, that is the death rates, are shown by coloring different counties of the US according to the death-rates' statistical information.

A user can easily locate which counties suffered the most due to cancer and which suffered the least by consulting the color scheme provided with this data map. Provided with the additional information related to the socioeconomic norms of a county one can investigate the reasons, hubs and links in the spread of the disease. This type of information can thus be very useful in healthcare and specially in epidemiology to prevent a potential outbreak of a viral disease.

Relational graphics:

In this type of visualization the variables can take any form or type. Like mentioned before, in these types of graphics the relation between two or more quantities is analyzed, which are not necessarily only time and space. An example of this kind of analysis can be number of deaths per million versus cigarette consumption pattern over a range of a spatial region.

The variable of time can also be added to this analysis, extending the experiment to extract the changing patterns of deaths because of cigarette consumption over different periods of times. The resulting graphic will show how effective the campaign, against smoking, really is over a period of time by observing the decrease in the deaths in different regions. If, for example, a few regions are showing resistance then the focus can be diverted to this particular area. More variables, like those related to sociopolitical norms, can further be added to pinpoint the troubles and ideally address the issues.

CONCLUSION

In today's age it is quite likely that big data will gain substantial potential and importance in order to shift the paradigm of the conventional humanitarian development process in almost every walk of life. It is, however, not a panacea to all the problems in the modern day world. Just like any other innovation, the wide scale adoption of big data is hindered by many potential challenges. In this section we discuss some of these challenges from two perspectives: technical and ethical. Correspondingly, we describe open issues and future work, which is required to address these challenges.

REFERENCES

[1] V. Mayer-Schonberger and K. Cukier (2013). Big Data: A Revolution that Will Transform how We Live, Work, and Think. Eamon Dolan/Houghton Mifflin Harcourt.

[2] James Manyika, Michael Chui, Peter Bisson, Jonathan Woetzel, Richard Dobbs, Jacques Bughin, Dan Aharon (2015). "The Internet of things: Mapping the value beyond the hype," tech. rep., McKinsey Global Institute, 06 2015.

[3] E. Siegel (2013). Predictive Analytics: The Power to Predict who Will Click, Buy, Lie, Or Die. John Wiley & Sons.

[4] L. A. Barroso, J. Clidaras, and U. Holzle (2013). "The datacenter as a computer: An introduction to the design of warehouse-scale machines," ¨ Synthesis lectures on computer architecture, vol. 8, no. 3, pp. 1–154.

[5] M. Hilbert (2013). "Big data for development: From information-to knowledge societies," Available at SSRN 2205145, 2013.

[6] "The home of the U.S. Government's open data." http://www.data.gov/. [Online; accessed 01-October-2015].

[7] L. Hoffmann (2012). "Data mining meets city hall," Communications of the ACM, vol. 55, no. 6, pp. 19–21.

[8] "The home of the U.K. Government's open data." https://www.data.gov.uk/. [Online; accessed 01-October-2015].

[9] C. Hartung, A. Lerer, Y. Anokwa, C. Tseng, W. Brunette, and G. Borriello (2010). "Open data kit: tools to build information services for developing regions," in Proceedings of the 4th ACM/IEEE International Conference on Information and Communication Technologies and Development, p. 18, ACM.

[10] J. Manyika (2013). Open data: Unlocking innovation and performance with liquid information. McKinsey.

[11] "Centro de Operaes Rio." http://centrodeoperacoes.rio/. [Online; accessed 05-October-2015].

[12] "Big data in action for development," tech. rep., The World Bank, 2014.

Corresponding Author

Parag Chandra Dutta*