APPLYING BIG DATA FOR MACHINE LEARNING PROCESS

Yew Kee Wong

School of Information Engineering, HuangHuai University, Henan, China.

ABSTRACT

In the information era, enormous amounts of data have become available on hand to decision makers. Big data refers to datasets that are not only big, but also high in variety and velocity, which makes them difficult to handle using traditional tools and techniques. Due to the rapid growth of such data, solutions need to be studied and provided in order to handle and extract value and knowledge from these datasets. Machine learning is a method of data analysis that automates analytical model building. It is a branch of artificial intelligence based on the idea that systems can learn from data, identify patterns and make decisions with minimal human intervention. Such minimal human intervention can be provided using big data analytics, which is the application of advanced analytics techniques on big data. This paper aims to analyse some of the different machine learning algorithms and methods which can be applied to big data analysis, as well as the opportunities provided by the application of big data analytics in various decision making domains.

KEYWORDS

Artificial Intelligence, Machine Learning, Big Data Analysis

1. INTRODUCTION

Resurging interest in machine learning is due to the same factors that have made data mining and Bayesian analysis more popular than ever. Things like growing volumes and varieties of available data, computational processing that is cheaper and more powerful, and affordable data storage. All of these things mean it's possible to quickly and automatically produce models that can analyse bigger, more complex data and deliver faster, more accurate results – even on a very large scale. And by building precise models, an organization has a better chance of identifying profitable opportunities – or avoiding unknown risks[1].

Because of new computing technologies, machine learning today is not like machine learning of the past. It was born from pattern recognition and the theory that computers can learn without being programmed to perform specific tasks; researchers interested in artificial intelligence wanted to see if computers could learn from data. The iterative aspect of machine learning is important because as models are exposed to new data, they are able to independently adapt. They learn from previous computations to produce reliable, repeatable decisions and results. It's a science that's not new – but one that has gained fresh momentum. While many machine learning algorithms have been around for a long time, the ability to automatically apply complex mathematical calculations to big data, over and over, faster and faster – is a recent development [2]. This paper will look at some of the different machine learning algorithms and methods which can be applied to big data analysis, as well as the opportunities provided by the application of big data analytics in various decision making domains.

2. HOW MACHINE LEARNING WORKS

Machine learning is a branch of artificial intelligence (AI) and computer science which focuses on the use of data and algorithms to imitate the way that humans learn, gradually improving its accuracy [3].

Machine learning is an important component of the growing field of data science. Through the use of statistical methods, algorithms are trained to make classifications or predictions, uncovering key insights within data mining projects. These insights subsequently drive decision making within applications and businesses, ideally impacting key growth metrics [4]. As big data continues to expand and grow, the market demand for data scientists will increase, requiring them to assist in the identification of the most relevant business questions and subsequently the data to answer them.

2.1. Machine Learning Algorithms

Machine learning algorithms can be categorize into three main parts:

1. **A Decision Process**: In general, machine learning algorithms are used to make a prediction or classification. Based on some input data, which can be labelled or unlabelled, your algorithm will produce an estimate about a pattern in the data.
2. **An Error Function**: An error function serves to evaluate the prediction of the model. If there are known examples, an error function can make a comparison to assess the accuracy of the model.
3. **A Model Optimization Process**: If the model can fit better to the data points in the training set, then weights are adjusted to reduce the discrepancy between the known example and the model estimate. The algorithm will repeat this evaluate and optimize process, updating weights autonomously until a threshold of accuracy has been met.

2.2. Types of Machine Learning Methods

Machine learning classifiers fall into three primary categories [5]:

Supervised machine learning
Supervised learning also known as supervised machine learning, is defined by its use of labelled datasets to train algorithms that to classify data or predict outcomes accurately. As input data is fed into the model, it adjusts its weights until the model has been fitted appropriately. This occurs as part of the cross validation process to ensure that the model avoids over fitting or under fitting. Supervised learning helps organizations solve for a variety of real-world problems at scale, such as classifying spam in a separate folder from your inbox. Some methods used in supervised learning include neural networks, naïve bayes, linear regression, logistic regression, random forest, support vector machine (SVM), and more.

Unsupervised machine learning
Unsupervised learning, also known as unsupervised machine learning, uses machine learning algorithms to analyse and cluster unlabelled datasets. These algorithms discover hidden patterns or data groupings without the need for human intervention. Its ability to discover similarities and differences in information make it the ideal solution for exploratory data analysis, cross-selling strategies, customer segmentation, image and pattern recognition. It's also used to reduce the number of features in a model through the process of dimensionality reduction; principal component analysis (PCA) and singular value decomposition (SVD) are two common approaches for this. Other algorithms used in unsupervised learning include neural networks, k-means clustering, probabilistic clustering methods, and more [6].

Semi-supervised learning
Semi-supervised learning offers a happy medium between supervised and unsupervised learning. During training, it uses a smaller labelled data set to guide classification and feature extraction from a larger, unlabelled data set. Semi-supervised learning can solve the problem of having not enough labelled data (or not being able to afford to label enough data) to train a supervised learning algorithm.

2.3. Practical Use of Machine Learning

Here are just a few examples of machine learning you might encounter every day [7]:

Speech Recognition: It is also known as automatic speech recognition (ASR), computer speech recognition, or speech-to-text, and it is a capability which uses natural language processing (NLP) to process human speech into a written format. Many mobile devices incorporate speech recognition into their systems to conduct voice search—e.g. Siri—or provide more accessibility around texting.

Customer Service: Online chatbots are replacing human agents along the customer journey. They answer frequently asked questions (FAQs) around topics, like shipping, or provide personalized advice, cross-selling products or suggesting sizes for users, changing the way we think about customer engagement across websites and social media platforms. Examples include messaging bots on e-commerce sites with virtual agents, messaging apps, such as Slack and Facebook Messenger, and tasks usually done by virtual assistants and voice assistants.

Computer Vision: This AI technology enables computers and systems to derive meaningful information from digital images, videos and other visual inputs, and based on those inputs, it can take action. This ability to provide recommendations distinguishes it from image recognition tasks. Powered by convolutional neural networks, computer vision has applications within photo tagging in social media, radiology imaging in healthcare, and self-driving cars within the automotive industry.

Recommendation Engines: Using past consumption behaviour data, AI algorithms can help to discover data trends that can be used to develop more effective cross-selling strategies. This is used to make relevant add-on recommendations to customers during the checkout process for online retailers.

Automated stock trading: Designed to optimize stock portfolios, AI-driven high-frequency trading platforms make thousands or even millions of trades per day without human intervention.

3. WHAT IS BIG DATA AND WHAT ARE ITS BENEFITS

Big data analytics has revolutionized the field of IT, enhancing and adding added advantage to organizations. It involves the use of analytics, new age tech like machine learning, mining, statistics and more. Big data can help organizations and teams to perform multiple operations on a single platform, store Tbs of data, pre-process it, analyse all the data, irrespective of the size and type, and visualize it too [8].

The Sources of Big Data:

Black Box Data
This is the data generated by airplanes, including jets and helicopters. Black box data includes flight crew voices, microphone recordings, and aircraft performance information.

Social Media Data
This is data developed by such social media sites as Twitter, Facebook, Instagram, Pinterest, and Google+.

Stock Exchange Data
This is data from stock exchanges about the share selling and buying decisions made by customers.

Power Grid Data
This is data from power grids. It holds information on particular nodes, such as usage information.

Transport Data
This includes possible capacity, vehicle model, availability, and distance covered by a vehicle.

Search Engine Data
This is one of the most significant sources of big data. Search engines have vast databases where they get their data.

The speed at which data is streamed, nowadays, is unprecedented, making it difficult to deal with it in a timely fashion. Smart metering, sensors, and RFID tags make it necessary to deal with data torrents in almost real-time. Most organizations are finding it difficult to react to data quickly. Not many years ago, having too much data was simply a storage issue [9]. However, with increased storage capacities and reduced storage costs are now focusing on how relevant data can create value.

There is a greater variety of data today than there was a few years ago. Data is broadly classified as structured data (relational data), semi-structured data (data in the form of XML sheets), and unstructured data (media logs and data in the form of PDF, Word, and Text files). Many companies have to grapple with governing, managing, and merging the different data varieties [10].

3.1. Advantages of Big Data

1. Today's consumer is very demanding. All customer wants to be treated as an individual and to be thanked after buying a product. With big data, supplier will get actionable data that they can use to engage with their customers one-on-one in real-time [11]. One way big data allows supplier to do this is that they will be able to check a complaining customer's profile in real-time and get info on the product(s) the customer is complaining about. Supplier will then be able to perform reputation management.

2. Big data allows supplier to re-develop the products/services they are selling. Information on what others think about their products, such as through unstructured social networking site text helps supplier in product development.

3. Big data allows supplier to test different variations of CAD (computer-aided design) images to determine how minor changes affect their process or product. This makes big data invaluable in the manufacturing process.

4. Predictive analysis will keep supplier ahead of their competitors. Big data can facilitate this by, as an example, scanning and analysing social media feeds and newspaper reports. Big data also helps supplier do health-tests on their customers, suppliers, and other stakeholders to help supplier reduce risks such as default.

5. Big data is helpful in keeping data safe. Big data tools help supplier map the data landscape of their company, which helps in the analysis of internal threats. As an example, supplier will know if their sensitive information has protection or not. A more specific example is that supplier will be able to flag the emailing or storage of 16 digit numbers (which could, potentially, be credit card numbers) [12].

6. Big data allows supplier to diversify their revenue streams. Analysing big data can give supplier trend-data that could help the supplier come up with a completely new revenue stream.

7. The supplier website needs to be dynamic if it is to compete favourably in the crowded online space. Analysis of big data helps supplier personalize the look/content and feel of their site to suit every visitor based on, for example, nationality and sex. An example of this is Amazon's IBCF (item-based collaborative filtering) that drives its "People you may know" and "Frequently bought together" features [13].

8. If the supplier is running a factory, big data is important because the supplier will not have to replace pieces of technology based on the number of months or years they have been in use. This is costly and impractical since different parts wear at different rates. Big data allows supplier to spot failing devices and will predict when the supplier should replace them.

9. Big data is important in the healthcare industry, which is one of the last few industries still stuck with a generalized, conventional approach. Big data allows a cancer patient to get medication that is developed based on his/her genes.

3.2. Challenging of Big Data

1. One of the issues with big data is the exponential growth of raw data. The data centres and databases store huge amounts of data, which is still rapidly growing. With the exponential growth of data, organizations often find it difficult to rightly store this data [14].

2. The next challenge is choosing the right big data tool. There are various big data tools, however choosing the wrong one can result in wasted effort, time and money too.

3. Next challenge of big data is securing it. Often organizations are too busy understanding and analysing the data, that they leave the data security for a later stage, and unprotected data ultimately becomes the breeding ground for the hackers.

4. CONCLUSIONS

So this study was concerned by understanding the interrelation between machine learning and big data analysis, what frameworks and systems that worked, and how machine learning can impact the big data analytic process whether by introducing new innovations that foster advanced machine learning process and escalating power consumption, security issues and replacing human in workplaces. The advanced big data analytics and machine learning algorithms with various applications show promising results in artificial intelligence development and further evaluation and research using machine learning are in progress.

REFERENCES

[1] Shi, Z., (2019). Cognitive Machine Learning. *International Journal of Intelligence Science*, 9, pp. 111-121.

[2] Lake, B.M., Salakhutdinov, R. and Tenenbaum, J.B., (2015). Human-Level Concept Learning through Probabilistic Program Induction. Science, 350, pp. 1332-1338.

[3] Silver, D., Huang, A., Maddison, C.J., et al., (2016). Mastering the Game of Go with Deep Neural Networks and Tree Search. Nature, 529, pp. 484-489.

[4] Fukushima, K., (1980). Neocognitron: A Self-Organizing Neural Network Model for a Mechanism of Pattern Recognition Unaffected by Shift in Position. Biological Cybernetics, 36, pp. 193-202.

[5] Lecun, Y., Bottou, L., Orr, G.B., et al., (1998). Efficient Backprop. Neural Networks Tricks of the Trade, 1524, pp. 9-50.

[6] McClelland, J.L., et al., (1995). Why There Are Complementary Learning Systems in the Hippocampus and Neocortex: Insights from the Successes and Failures of Connectionist Models of Learning and Memory. Psychological Review, 102, pp. 419-457.

[7] Kumaran, D., Hassabis, D. and McClelland, J.L., (2016). What Learning Systems Do Intelligent Agents Need? Complementary Learning Systems Theory Updated. Trends in Cognitive Sciences, 20, pp. 512-534.

[8] S. Del. Rio, V. Lopez, J. M. Bentez and F. Herrera, (2014). On the use of mapreduce for imbalanced big data using random forest, Information Sciences, 285, pp. 112-137.

[9] MH. Kuo, T. Sahama, A. W. Kushniruk, E. M. Borycki and D. K. Grunwell, (2014). Health big data analytics: current perspectives, challenges and potential solutions, International Journal of Big Data Intelligence, 1, pp. 114-126.

[10] R. Nambiar, A. Sethi, R. Bhardwaj and R. Vargheese, (2013). A look at challenges and opportunities of big data analytics in healthcare, IEEE International Conference on Big Data, pp. 17-22.

[11] Z. Huang, (1997). A fast clustering algorithm to cluster very large categorical data sets in data mining, SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery.

[12] M. D. Assuno, R. N. Calheiros, S. Bianchi, M. a. S. Netto and R. Buyya, (2015). Big data computing and clouds: Trends and future directions, Journal of Parallel and Distributed Computing, 79, pp. 3-15.

[13] I. A. T. Hashem, I. Yaqoob, N. Badrul Anuar, S. Mokhtar, A. Gani and S. Ullah Khan, (2014). The rise of big data on cloud computing: Review and open research issues, Information Systems, 47, pp. 98-115.

[14] L. Wang and J. Shen, (2013). Bioinspired cost-effective access to big data, International Symposium for Next Generation Infrastructure, pp. 1-7.

Author

Prof. Yew Kee Wong (Eric) is a Professor of Artificial Intelligence (AI) & Advanced Learning Technology at the HuangHuai University in Henan, China. He obtained his BSc (Hons) undergraduate degree in Computing Systems and a Ph.D. in AI from The Nottingham Trent University in Nottingham, U.K. He was the Senior Programme Director at The University of Hong Kong (HKU) from 2001 to 2016. Prior to joining the education sector, he has worked in international technology companies, Hewlett-Packard (HP) and Unisys as an AI consultant. His research interests include AI, online learning, big data analytics, machine learning, Internet of Things (IOT) and blockchain technology.