Machine Learning in the Age of Automation

Cesar Cuevas and Israel B. Cuevas
Andrew L. Mackey (Advisor)
Computer and Information Sciences Department
University of Arkansas - Fort Smith
Fort Smith, Arkansas 72913–3649

***Abstract*—Technological advancements continue to push new boundaries and improve the lives of numerous people across a wide range of industries. Many of these technological advancements have been made possible through automation and the field of machine learning. Automobiles now have the ability to move between locations autonomously with little to no input from consumers. Machine learning has also managed to facilitate the discovery of new effective drug treatments to save the lives of many while lowering the costs of achieving these goals. From self-driving cars and new drug discovery to manufacturing along with many others, advancements in the field of machine learning enable the completion of tasks once thought to be existent only by humans. The goal of this presentation is to survey the advancements in the field of machine learning and convey how data and intelligent algorithms will affect the automation of tasks in the future.**

I. Introduction

A new era of computing that draws comparisons to the Industrial Revolution has started in which organizations are able to leverage machine learning in new ways to improve the lives of many individuals. Evidence of this can be seen in a variety of different industries as organizations are quickly moving to exploit the hidden trends within the data they are generating [1]. The emergence of new technology, such as smartphones and wearable devices, offers new ways to generate and provide data previously unavailable to those who need it most (i.e. doctors, medical professionals, etc.). As the computational power, connectivity, volume of data, and algorithms to extract meaningful trends continue to increase, machine learning efforts will have a profound impact on the world's operational processes through automation.

II. Understanding Machine Learning

Machine Learning is a field of study that combines computer algorithms with data to yield predictive models capable of facilitating analysis with unknown targets. Typical computer applications are generally comprised of a distinct set of instructions that perform a specific task as determined by the developer. These sets of instructions or operations, referred to as algorithms, generally solve problems by explicitly coding the methods to achieve a solution. In contrast, machine learning requires that computers be adaptive in their approach to generating output for a specified input data set. Furthermore, it also requires that computers must have the ability to modify their actions to improve the accuracy of the predictive models built by the algorithm.

The algorithms leveraged by machine learning applications require data in order to formulate a model that has predictive capabilities. While the study of data is inherent across multiple disciplines, machine learning attempts to borrow theorems from various domains of research, such as statistics and biology, to generate algorithms that can effectively and efficiently extract information from the data provided. As data sets increase in both size and complexity, machine learning attempts to ascertain trends autonomously where the feasibility of the same computation through human labor would be questionable, if not impractical.

Many organizations and CIO's have identified data as the next "natural resource" [2]. The advent of many diverse data-generating technologies have paved the way for new avenues of obtaining sources of valuable information for use with machine learning applications. One primary example of a device that has accelerated the generation of data is the smartphone. In 2015, the adoption rate for these devices was as high as 68 in the United States of America [3]. Unlike their relatively stationary counterparts, the desktop and laptop computers, smartphones brought a new form of connected mobility to consumers that is substantially more difficult to achieve in other form factors. This enabled consumers to leverage a new form of web searching, online shopping, communication, and access many other services through mobile devices. In many cases, each click, web search, online purchase or other interaction with a system typically produces a transaction of data that is temporal in nature. With the appropriate machine learning algorithms, organizations can leverage this data to improve their business processes or even find new ways to automate tasks to achieve higher levels of efficiency.

One key difference between the transactions occurring on mobile devices and laptop/desktop computers is the frequency with which they occur [4]. This was a contributing factor to the present era of "big data" – a term promulgated to generically describe the volume, velocity and varieties of data being generated. As platforms emerged to tackle the issues surrounding big data accompanied by the declining prices of storage, data became ubiquitous and voluminous – a prerequisite component for machine learning algorithms to be successful. These systems, which are used to manage this vast amount of data, introduced several new programming paradigms and algorithms for implementing machine learning at scale without having performance inhibited by the factors of big data. As the prices of storage continue to diminish, the total cost of ownership for the implementation of systems to perform machine learning at scale is also in decline as a number of platforms, such as Hadoop or Apache Spark, leverage commodity hardware that is generally affordable in comparison to the hardware needed to power highly active

transaction processing systems. This, along with the existence of platforms capable of managing large amounts of data and implementing various machine learning algorithms, has lowered the barriers to entry into the field of machine learning.

III. Automation and Society

The rise of machine learning and automation brought many benefits transcending various industries, including manufacturing, transportation, utilities, etc. One prominent example of machine learning and automation can be found in the transportation industry through the use of self-driving cars, such as those provided by Tesla or Waymo [5] [6]. In other cases, the automation of tasks through machine learning can facilitate the solving of complex problems while requiring little to no human intervention [7]. Results from this have yielded lower costs and improvements in quality assurance and productivity levels. The development of new computing technologies continues to perpetuate growth in automation and machine learning endeavors.

As machine learning and automation becomes more mainstream, anxiety and stress continue to build with the general population as the threat of job loss continues to be associated with the term automation. A reduction of positions involving repetitive tasks is imminent when the cost and feasibility of automation exceed the cost of human labor. Some research contends that these positions are at the greatest risk when organizations attempt to maximize the productivity of its labor force while reducing the costs associated with the operations of the firm [8]. During the recession that transpired in 2008, many employers underwent a reduction in headcount of employees and automated many of these positions. As the economic markets showed improvement, many employers did not restore these automated positions to their previous workers. Automation can present many challenges in the future. However, other factors (i.e. governmental restrictions, tax policy, etc.) may be serving as impediments to job growth [8].

IV. Machine Learning in the Medical Field

The medical field has numerous applications in which automation efforts can improve the quality of care patients received. One such example that applies to patients receiving dental care demonstrates an effort to automate time-consuming processes by determining which individuals are candidates for tooth cusps based on geometric features of teeth [9]. By leveraging Artificial Neural Networks (ANN) and the back-propagation algorithm for the calculation of weights, researchers were able to analyze a set of data to perform a method of supervised learning known as classification. The research was able to yield a model that would correctly classify between 93.3% – 93.5% of the data – a level of precision the authors note would be sufficient for exploration within clinical practices [9].

Another case in which automation and machine learning demonstrate substantial potential would be in the field of echocardiography. This area is responsible for the digital images of high spatiotemporal resolution in which analysis commences involving dimensions, volumes, wall thickness/mass, and wall motion [10]. At the conclusion of each examination, a significant amount of data is generated for analysis. In one study, the author notes that the process of involving echocardiographic images is highly inefficient and time-consuming. Given the sheer volume of data, cognitive overload is a concern as the potential for costly diagnostic errors increases. As a result, one study involving patients was able to produce results that paralleled conventional methods [10].

V. Machine Learning at Scale

One of the goals with the research presented here was to establish a cluster of systems using commodity hardware, large data sets comprised of 100,000 - 100+ million records with multiple attributes (e.g. age, zip code, etc.), and machine learning algorithms, such as C4.5 decision trees, k-means, k-nearest neighbor, linear and logistic regression. A total of 50 individual systems equipped with 16GB of memory and 100+ GB of secondary storage were dedicated to this project. Each system was running OpenSuSE 42.2 with Hadoop and Apache Spark.

To demonstrate the efficiency of machine learning at scale through the utilization of parallel processing, linear regression was implemented on the cluster of systems with the data provided. While this form of regression analysis is common within various fields, including statistics and economics, this research attempts to demonstrate how machine learning models can be built to extract or generalize patterns. Linear regression allows practitioners to review the relationships between attributes within a set of data [11]. When the relationship between a set of data needs to be plotted (i.e. sales by month), it may be possible to build a linear equation through the data by minimizing the sum of the squared residuals, which is the a line that minimizes the difference between the observed and fitted values.

The following equations can be used to calculate the mean of *x* and *y*:

$$\bar{x} = \frac{1}{n}\sum_{i}^{n} x_i \qquad \bar{y} = \frac{1}{n}\sum_{i}^{n} y_i$$

A standard linear equation can be represented as $y = m(x)+b$, but the work here will represent the slope as the variable b and the y-intercept as the variable a. The variable $\hat{y}$ represents the *predicted value* for *y* in the linear regression equation:

$$\hat{y} = a + bx$$

Slope (represented as *b*) will be calculated for the equation using the following equation:

$$b = \frac{\sum_{i}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i}^{n}(x_i - \bar{x})^2}$$

The intercept (represented as *a*) is the difference between the mean of *y* (represented as $\bar{y}$) and product of the slope and the mean of *x* (represented as $\bar{x}$):

$$a = \bar{y} - b\,\bar{x}$$

On smaller data sets, this approach performs reasonably well with minimal obstacles on individual machines. However, as the volume and velocity of the data increases, the scalability and performance of linear regression algorithms can be easily impeded. For some of the machine learning algorithms implemented for this study, a linear increase in the amount of data did not result in a linear increase in time. To curtail the impediments of machine learning algorithms given the aforementioned issues associated with big data, the parallelization of machine learning processes may be necessary to achieve reasonable levels of performance. This becomes more critical in algorithms requiring multiple iterations through the data to achieve the best results when building machine learning models.

VI. Conclusion

As data continues to grow at unprecedented rates, new ways to utilize machine learning will continue to emerge. Various industries will undergo significant changes as repetitive tasks are replaced through automation efforts to improve efficiency and productivity. In many cases, this has the potential to optimize the processes that currently experience substantial costs and delays, such as those observed in the field of health care. The algorithms used to extract insights from the data will continue to evolve while improving the accuracy of the models produced. As academic research and industry continue to expand the field of machine learning, new challenges will be inevitable as integration of machine learning becomes more mainstream within society.

Acknowledgment

The authors would like to thank our mentor, Andrew L. Mackey, for his unconditional support and contributions to this research. In the words of Shawn Hitchcock, "A mentor empowers a person to see a possible future, and believe that it can be obtained."

References

[1] T. Singer, "Researchers use Twitter to Track the Flu in Real Time," *News@NorthEastern,* 05-May-2017. [Online]. Available: http://news.northeastern.edu/2017/05/researchers-use-twitter-to-track-the-flu-in-real-time/. [Accessed: 22-March-2017].

[2] B. Picciano, "Why Big Data is The New Natural Resource," *Forbes*, 03-June-2014. [Online]. Available: https://www.forbes.com/sites/ibm/2014/06/30/why-big-data-is-the-new-natural-resource/#2a6198cc6628. [Accessed: 28-April-2017].

[3] M. Anderson, "Technology Device Ownership: 2015," *Pew Research Center: Internet, Science & Tech*, 29-Oct-2014. [Online]. Available: http://www.pewinternet.org/2015/10/29/technology-device-ownership-2015/. [Accessed: 22-March-2017].

[4] L. Eadicicco, "More People Now Shop on Amazon Using Smartphones and Tablets Than Computers," *Time*, 28-Dec-2015. [Online]. Available: http://time.com/4162188/amazon-holiday-shopping-statistics-2015/. [Accessed: 24-March-2017].

[5] Tesla, "Tesla", 2017. [Online]. Available: https://www.tesla.com. [Accessed: 30-April-2017].

[6] Waymo, "Waymo", 2017. [Online]. Available: https://www.waymo.com. [Accessed: 30-April-2017].

[7] D. T. Pham, A. A. Afify, "Machine-learning techniques and their applications in manufacturing," *Proceedings of the Institution of Mechanical Engineers,Journal of Engineering Manufacture,* vol. 219, no. 5, pp. 395-412.

[8] M. Castells, *The Rise of the Network Society*, 2nd ed. U.S: Wiley-Blackwell, 2009.

[9] S. Raith, P. E. Vogel, N. Anees, C. Keul, J. -F.Guth, D. Edelhoff and H. Fischer, "Artificial Neural Networks as a powerful numerical tool to classify specific features of a tooth based on 3D scan data," *Computers in Biology and Medicine,* vol. 80,pp. 65-76.

[10] A. J Tajik, "Machine Learning for Echocardiographic Imaging," *Journal of the American College of Cardiology,* vol. 68, no. 21, 2016, pp. 2296-2298.

[11] R. A. Johnson, I. Miller, and J. E. Freund, *Miller & Freund's Probability and Statistics for Engineers*, 8th ed. University of Wisconsin-Madison, 2010.