Scientific Journal of Impact Factor (SJIF): 4.72

e-ISSN (O): 2348-4470
p-ISSN (P): 2348-6406

International Journal of Advance Engineering and Research Development

Emerging Trends and Innovations in Electronics and Communication Engineering - ETIECE-2017

Volume 5, Special Issue 01, Jan.-2018 (UGC Approved)

Machine Learning for Data Analysis and its Applications

Firasath Nabi[1], Sanjay Jamwal[2], Kumar Padmanbh[3]

[1]Department of Computer Sciences, BGSBU, Rajouri, J&K
[2]Department of Computer Sciences, BGSBU, Rajouri, J&K
[3]Department of Computer Science & Eng., LNMIIT, Jaipur, Rajasthan

Abstract — *Gathering and maintaining large collections of data is one thing, but extracting useful information from these collections is even more challenging. Most industries working with large amounts of data have recognized the value of machine learning technology. By gleaning insights from this data – often in real time – organizations are able to work more efficiently or gain an advantage over competitors. The interest in predicting future outcomes using your data? Machine learning is the process of developing, testing, and applying predictive algorithms to achieve this goal.*
This paper focusses on the available machine learning approach for real-time processing of data. The paper will also present a brief review few applications of machine leaning in real time prediction, think of business cases such as product recommendation, market forecasting, segmentation of customers, fraud detection or churn prevention. Machine learning techniques can solve such applications using a set of generic methods that differ from more traditional statistical techniques. The emphasis is on real-time and highly scalable predictive analytics, using fully automatic and generic methods that simplify some of the typical data scientist tasks.

Keywords*-Analytics; Data; Machine Learning; Real-time; Sensor Data;*

I. INTRODUCTION

The increasing interest in data mining, or the use of historical data to discover regularities and improve future decisions, follows from the confluence of several recent trends: the falling cost of large data storage devices and the increasing ease of collecting data over networks; the development of robust and efficient machine learning algorithms to process this data; and the falling cost of computational power, enabling use of computationally intensive methods for data analysis. The field of machine learning, has already produced practical applications in such areas as analyzing medical outcomes, detecting credit card fraud, predicting customer purchase behavior, predicting the personal interests of Web users, and optimizing manufacturing processes. It has also led to a set of fascinating scientific questions about how computers might automatically learn from past experience [1]. Machine Learning (ML) is the science of how computers can improve their perception, cognition, and action with experience; in other words, ML teaches computers to learn from experience by using adaptive algorithms combined with computational methods to learn information directly from data without relying on code as a model as shown in table 1. Machine Learning is about how computers can act by themselves without being explicitly programmed. As a field of Artificial Intelligence (AI), ML improves computers' performance from data, knowledge, experience, and interaction. Machine Learning started with two breakthroughs [1, 2]:

1. The Arthur Samuel's pioneering work on computer gaming and AI, that made possible computers to learn from themselves instead of instructing them everything they need to know and how to do tasks.

2. The Internet growth of the past decade making available huge amounts of digital information for analysis.

Widely publicized examples of machine learning applications you may be familiar with [3]:

- The heavily hyped, self-driving Google car? Development of algorithms for auto-piloting drones. The essence of machine learning.
- Online recommendation offers such as those from Amazon and Netflix? Machine learning applications for everyday life. Knowing what customers are saying about you on Twitter? Machine learning combined with linguistic rule creation.
- Actuarial estimates of financial damage of storms and natural disasters, prediction of popular election outcomes.
- Pattern Recognition: Face, Character or Speech Recognition, Association Analysis.
- Fraud detection? One of the more obvious, important uses in our world today.

Engineers realized that it was far more efficient to code computers to think and understand the world like humans, giving access to all the information available on the internet and letting them to learn; keeping the innate advantages computers hold over humans: speed, accuracy, or lack of bias.

	General Analysis Techniques	Machine Learning Techniques
Who creates Model	Human	Machine (program)
Examples of Model	Linear model, Polynomial Model, etc.	Models that cannot be understood intuitively (e.g., neural model)

Table 1. Analyzing Techniques

II. EVOLUTION OF MACHINE LEARNING

or a similar sans-serif font). 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 [4,5]. 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's gaining 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 [6].

III. WHAT IS DATA ANALYTICS?

Analytics is an encompassing and multidimensional field that uses mathematics, statistics, predictive modeling and machine-learning techniques to find meaningful patterns and knowledge in recorded data, increasingly with the aid of specialized systems and software. Today, we add powerful computers to the mix for storing increasing amounts of data and running sophisticated software algorithms – producing the fast insights needed to make fact-based decisions. By putting the science of numbers, data and analytical discovery to work, we can find out if what we think or believe is really true. And produce answers to questions we never thought to ask. Analysis of data can uncover correlations and patterns. There's less need to rely on guesses or intuition. Analysis of data can uncover correlations and patterns. There's less need to rely on guesses or intuition. And it can help answer the following types of questions [7]:

What happened?
How or why did it happen?
What's happening now?
What is likely to happen next?

There are three predominant types of analytics in use today are descriptive analysis (What happened?), predictive analysis (What will happen?) and perspective analysis (How can we make it happen?).

The work of a data analyst lies in inference, which is the process of deriving conclusions that are solely based on what the researcher already knows; for example, running through a number of data sets to look for meaningful correlations between each other. Data analytics technologies and techniques are widely used in: Commercial industries to enable organizations to make more-informed business decisions and by scientists and researchers to verify or disprove scientific models, theories and hypotheses [8].

3.1. Data Analysis Process

When you have a good understanding of a phenomenon, it is possible to make predictions about it. Data analysis helps us to make this possible through exploring the past and creating predictive models.
The data analysis process is composed of the following steps [8, 9, 10]:

- **The statement of problem**
- **Obtain your data**
- **Clean the data**
- **Normalize the data**
- **Transform the data**
- **Exploratory statistics**
- **Exploratory visualization**
- **Predictive modeling**
- **Validate your model**
- **Visualize and interpret your results**
- **Deploy your solution**

1. **The problem**
   The problem definition starts with high-level questions such as how to track differences in behavior between groups of customers, or what's going to be the gold price in the next month.
2. **Data preparation**
   How to obtain, clean, normalize, and transform the data into an optimal dataset, trying to avoid any possible data quality issues such as invalid, ambiguous, out-of-range, or missing values. This process can take a lot of time. Analyzing data that has not been carefully prepared can lead you to highly misleading results.
3. **Data exploration:** Data exploration is essentially looking at the data in a graphical or statistical form trying to find patterns, connections, and relations in the data.
4. **Visualization:** is used to provide overviews in which meaningful patterns may be found.
5. **Predictive modeling:** Create or choose a statistical model trying to best predict the probability of an outcome. The model evaluation helps us to ensure that our analysis is not over-optimistic or over-fitted. Two different ways to validate the model:

- Cross-validation: We divide the data into subsets of equal size and test the predictive model in order to estimate how it is going to perform in practice
  - Hold-Out: Mostly, large dataset is randomly divided in to three subsets: training set, validation set, and test set.

6. **Visualization of Results:** This is the final step in our analysis process and we need to answer the following questions:
   How is it going to present the results?
   Where is it going to be deployed?
7. **Data visualization:** an important part of our data analysis process because it is a fast and easy way to do an exploratory data analysis through summarizing their main characteristics with a visual graph, tabular reports, 2D plots, dashboards, or infographics.

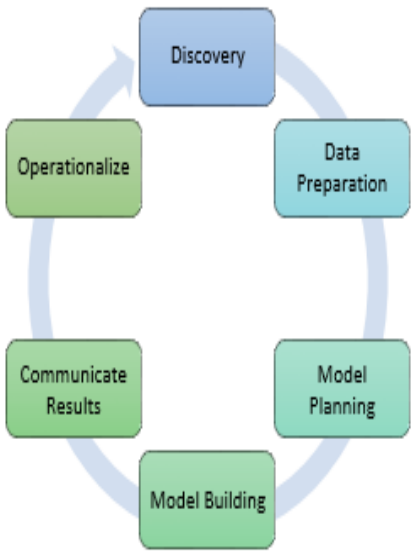

Figure 2. Data Analytic Life Cycle

Supervised Learning Algorithm	
Nearest Neighbor	Classification
Naive Bayes	Classification
Decision Trees	Classification
Classification Rule Learners	Classification
Linear Regression	Numeric prediction
Regression Trees	Numeric prediction
Model Trees	Numeric prediction
Neural Networks	Dual use
Support Vector Machines	Dual use
Unsupervised Learning Algorithms	
Association Rules	Pattern detection
k-means clustering	Clustering
Meta-Learning Algorithms	
Bagging	Dual use
Boosting	Dual use
Random Forests	Dual use

Figure 3. Types of Machine Leaning Approaches for Data Analytics

IV. MACHINE LEARNING APPROACHES

Understanding the categories of learning algorithms is an essential first step towards to analyze the data to drive the desired action. To apply the learning process to real-world tasks, any machine learning algorithm can be deployed by following these steps [15]:

1. Data collection
2. Data exploration and preparation
3. Model training
4. Model evaluation
5. Model improvement

Depending on the type of analysis to be performed on the huge data, Machine learning algorithms are divided into categories according to their analytic purpose [07, 11, 12] as shown in figure 3.

1.Predictive Modelling

A predictive model is used for tasks that involve, the prediction of one value using other values in the dataset. The learning algorithm attempts to discover and model the relationship between the target feature (the feature being predicted) and the other features. The process of training a predictive model is known as supervised learning. The target

values provide a way for the learner to know how well it has learned the desired task. Because predictive models are given clear instruction on what they need to learn and how they are intended to learn it, the process of training a predictive model is known as supervised learning.
The often-used machine learning task for predictive analysis is known as classification. Such Machine Learning models find uses as following: An e-mail message is spam, A person has cancer, A football team will win or lose, An applicant for a bank loan [13].
Due to the wide use of classification models in machine learning, there are many types of classification algorithms, with strengths and weaknesses suited for different types of input data. Supervised learners can be used to predict numeric data such as income, laboratory values, test scores, or counts of items. Regression methods are widely used for forecasting.

2.Descriptive Modelling

A descriptive model is used for tasks that would benefit from the insight gained from summarizing data in new and interesting ways. In a descriptive model, there is no target to learn, the process of training a descriptive model is called unsupervised learning. The descriptive modeling task called pattern discovery is used to identify useful associations within data. Pattern discovery is often used for market basket analysis on retailers' transactional purchase data. Here, the goal is to identify items that are frequently purchased together, such that the learned information can be used to refine marketing tactics [14].
The descriptive modeling task of dividing a dataset into homogeneous groups is called clustering. This is sometimes used for segmentation analysis that identifies groups of individuals with similar behavior or demographic information, so that advertising campaigns could be tailored for particular audiences.

3.Meta- Learning Modelling

A meta-learning algorithm uses the result of some learnings to inform additional learning. It is not tied to a specific learning task, but is rather focused on learning how to learn more effectively. This can be beneficial for very challenging problems or when a predictive algorithm's performance needs to be as accurate as possible.

4.1. Why Machine Learning?

This field originated in an environment where available data, statistical methods, and computing power rapidly and simultaneously evolved. Traditionally, data science has always been dominated by trial-and-error analysis, an approach that becomes impossible when datasets are large and heterogeneous. Growth in data necessitated additional computing power, which in turn spurred the development of statistical methods to analyze large datasets. Ironically, availability of more data usually leads to fewer options in constructing predictive models, because very few tools allow for processing large datasets in a reasonable amount of time [15]. Machine learning methods are particularly effective in situations where deep and predictive insights need to be uncovered from data sets that are large, diverse and fast changing — Big Data. Machine learning easily outperforms traditional methods on accuracy, scale, and speed [16,17]. The scalability not only allows predictive solutions based on sophisticated algorithms to be more accurate, it also drives the importance of software's speed to interpret the billions of rows and columns in real-time and to analyze live streaming data. The more data you provide to the system, the more it can learn from it, returning all of the clues you were looking for. Most industries working with large amounts of data have recognized the value of machine learning technology in the following [18, 19,20,21]:

- **Financial services**

Banks and other businesses use machine learning technology for two key purposes:
To identify important insights in data, and prevent fraud.
The insights can identify investment opportunities, or help investors know when to trade.

- **Government**

Government agencies such as public safety and utilities have a particular need for machine learning since they have multiple sources of data that can be mined for insights. Analyzing sensor data, for example, identifies ways to increase efficiency and save money and minimize identity theft.

- **Oil and gas**

The number of machine learning use cases for this industry is vast – and still expanding. Finding new energy sources and analyzing minerals in the ground, predicting refinery sensor failure and streamlining oil distribution to make it more efficient and cost effective.

- **Transportation**

Analyzing data to identify patterns and trends, which relies on making routes more efficient and predicting potential problems to increase profitability. The data analysis and modeling aspects of machine learning are important tools to delivery companies, public transportation and other transportation organizations.

- **Health Care**

Machine learning is a fast-growing due to the advent of wearable devices and sensors that can use data to assess a patient's health in real time. The technology can also help medical experts analyze data to identify trends that may lead to improved diagnoses and treatment .

- **Marketing and Sales**

Websites recommending items you might like based on previous purchases are using machine learning to analyze your buying history – and promote other items you'd be interested in. This ability to capture data, analyze it and use it to personalize a shopping experience (or implement a marketing campaign) is the future of retail.

V. CONCLUSION

Machine learning is most successful when it augments rather than replaces the specialized knowledge of a subject-matter expert. Although machine learning is used widely and has tremendous potential, it is important to understand its limits. Machine learning, at this time, is not in any way a substitute for a human brain. It has very little flexibility to extrapolate outside of the strict parameters it learned and knows no common sense. With this in mind, one should be extremely careful to recognize exactly what the algorithm has learned before setting it loose in the real-world settings.

REFERENCES

[1] https://www.sas.com/en_us/insights/analytics/machine-learning.html
[2] Alexandra Moraru, Marko Pesko," Using Machine Learning on Sensor Data", Journal of Computing and Information Technology - CIT 18, 2010, 4, 341–347, doi:10.2498/cit.1001913, 2010.
[3] Tom M. Mitchell, "Machine Learning and Data Mining", Communications of the ACM November 1999/Vol. 42, No. 11, 1999.
[4] Brett Lantz, "Machine Learning with R", Chapter 1, 2nd ed, Packt Publishing Company, New Delhi, 2015.
[5] http://searchdatamanagement.techtarget.com/definition/data-analytics
[6] Go Takami, Moe Tokuma,"Machine Learning Applied to Sensor Analysis",Yokogawa Technical Report English Edition Vol.59 No. 1, 2016.
[7] Hector Cuesta, "Practical Data Analytics", Chapter 1, 2nd ed, Packt Publishing Company, New Delhi, 2015.
[8] https://www.sas.com/en_us/insights/analytics/what-is-analytics.html
[9] SB Kotsiantis, I Zaharakis, P Pintelas, "Supervised machine learning: A review of classification techniques", 2007.
[10] Daniel Gillblad , "On practical machine learning and data analysis", Doctoral Thesis, Stockholm, Sweden 2008.
[11] https://www.sas.com/en_us/insights/analytics/predictive-analytics.html
[12] Matt K. Smith, Charles C. Castello, Joshua R. New, "Machine Learning Techniques Applied ro Sensor Data Correction In Building Technologies", 12th International Conference on Machine Learning and Applications, 2013.
[13] Norbert Krupa, " Applying Machine Learning to IoT Sensors", https://www.talend.com/category/blog/, 2017.
[14] https://www.r-bloggers.com/how-to-use-data-analysis-for-machine-learning-example-part-1/
[15] R Burbidge, M Trotter, B Buxton, S Holden, "Drug design by machine learning: support vector machines for pharmaceutical data analysis", Computers & chemistry, - Elsevier, 2001.
[16] Iza Moise, Evangelos Pournaras, Dirk Helbing, "Introduction to Data Mining and Machine Learning Techniques",
[17] F Sebastiani, "Machine learning in automated text categorization", ACM computing surveys (CSUR), 2002.
[18] H Nielsen, S Brunak, G von Heijne, "Machine learning approaches for the prediction of signal peptides and other protein sorting signals", Protein engineering, 1999.
[19] Michael Walker, "Machine Learning Applications in the Real World", Rose Buisness Technologies.
[20] Christine PreisachHans BurkhardtLars Schmidt-ThiemeReinhold Decker, "Data Analysis, Machine Learning and Applications", Proceedings of the 31st Annual Conference of the Gesellschaft für Klassifikation e.V., Albert-Ludwigs-Universität Freiburg, March 7–9, 2007.
[21] Myra SpiliopoulouLars Schmidt-ThiemeRuth Janning, "Data Analysis, Machine Learning and Knowledge Discovery", 2014.

