ISSN: 2319-8753

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 8, August 2014

Big Data Analytics using Meta Machine Learning

Bharati Suvalka[1] ,Sarika kandelwal[2] , Sidharth Singh Sisodia[3]

M.Tech, Dept of CSE, GITS Rajasthan Technical University, Kota, Rajasthan, India[1].

Associate Professor, GITS Rajasthan Technical University, Kota, Rajasthan, India[2].

Assistant Professor, Dept of CSE, PIT Pacific University, Rajasthan, India[3].

ABSTRACT: As "Big Data" grows bigger at a rapid speed, Machine Learning (MACHINE LEARNING) techniques have come to play a vital role in automatic data processing and analytics across a wide spectrum of application domains. However, lack of well-defined values in choosing MACHINE LEARNING algorithms suitable for a given problem remains a major challenge. Today this choice depends mainly upon empirical rules such as the size of training data, number of dissimilar labels, need for interpretable decision boundaries, and real-time memory constraint. It is often also guided by realistic factors such as readily available code and comfort level of the programmers, and empirically unwavering parameters finely tuned by repeated experiments. We propose to lay the foundations of the next generation of domain agnostic MACHINE LEARNING techniques which will be able to sum up "a priori" knowledge of MACHINE LEARNING successes across domains, in analytics framework. Our goal is to alter the difficult alchemy involved in using MACHINE LEARNING techniques that take years to master into a simple skill that can be readily adapted by practitioners across fields. We wish to refer to this science as "Meta-Machine Learning" (MMACHINE LEARNING).

I.INTRODUCTION

Given the huge data explosion, largely due to the pervasiveness of the Internet, there is an urgent need for automate large-scale data analytics. Big data analytics involve processing heterogeneous data from various distributed data source producing complete data set[1].Big Data technologies describe a new generation of technologies and architectures, designed so organizations can economically extract value from very large volumes of a wide variety of data by enabling high-velocity capture, discovery, and/or analysis.[2] Advances in MACHINE LEARNING thus far, have addressed this need by developing efficient statistical techniques that learn from data. These include technique such as 'Supervised Learning' and 'Unsupervised Learning' designed to solve tasks such as data classification and clustering that arise across diverse fields. Given their broad based application, these MACHINE LEARNING techniques appeal to a wide audience in computing. Researchers in Computer Security use them to detect anomalous behavior in streaming data.

In the area of Green Computing and Smart Energy, MACHINE LEARNING techniques are used to learn energy usage patterns and matching them with real-time demand response. Computational Biologists are using MACHINE LEARNING for time-series data models to unravel the mysteries of human genome. Complex graphical models are being used by Linguists to discover syntactic patterns in written languages and words. Even outside Computer Science, kernel process and Bayesian predictions have assisted financial analysts to profitable proprietary trading tactics, and astronomers to cluster stars. Most of these applications have trust on years of manual knowledge transfer between developments, and several empirical refinements to the models, to make MACHINE LEARNING techniques work in any particular domain.

The lack of scalability across discipline, the need for manual knowledge and expertise transfer, and the burdensome practice of fine tuning "magic" parameters represent the main issues that have held back MACHINE LEARNING from realizing its full impending. We suppose that formalizing and addressing these issues will lead to an unprecedented

DOI: 10.15680/IJIRSET.2014.0308072

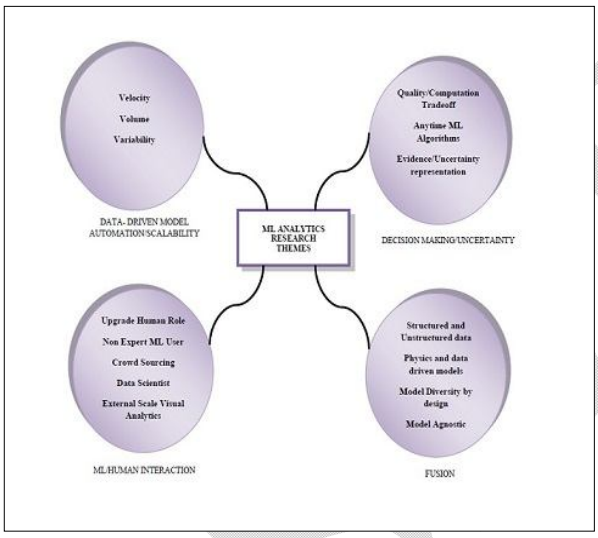
ISSN: 2319-8753

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 8, August 2014

deployment of MACHINE LEARNING techniques in more diverse fields than yet before. The goal of the planned workshop is to formalize the development of the Meta-Machine Learning framework that can be a "plug and play" tool in the hands of practitioners across a wide range of disciplines and applications.



(a) ML analytics on Big Data

II.IMPLEMENTATION OF MACHINE LEARNING ON BIG DATA

"Machine learning" generally adheres to the same principles as human learning, but usually and explicitly within a particular problem domain and with foundational mathematical rigor rarely present in the mental picture we have of human learning [3]. Today the MACHINE LEARNING community is focused on finding the best models of data for analytics tasks such as categorization and clustering. The application constraints are understood and a trial-and-error based methodology is adopted to find a suitable data model for the task. Existing MACHINE LEARNING techniques are refined or adapted based on the domain requirements. The field of machine learning is incredibly rich and diverse, but from the vast literature, we have identified two classes of techniques that are particularly amenable to large-scale machine learning. The first is stochastic gradient descent, representative of online learners that can easily scale to large datasets. The second is ensemble methods, which allow us to parallelize training in an nearly embarrassingly parallel manner, yet retain high levels of effectiveness. Both are well known in the machine learning literature, and together they form a powerful combination. These techniques occupy the focus of our implementation efforts.[4]

"Machine learning" generally adheres to the same principles as human learning, but usually and explicitly within a particular problem domain and with foundational mathematical rigor rarely present in the mental picture we have of human learning.[5] For example let us consider the domain of smart energy. The job is to find out the pattern of energy usage and model it efficiently to make available a real-time demand response system. A practitioner charged with the task would go about understanding the working of a Hidden Markov Model (HMM) and perform several experiments to tune it to specific needs. There are many problems to this "ad-hoc" MACHINE LEARNING application: (i) it does not guarantee best results because the choice of HMM (or other techniques such as Neural Networks) depends on the practitioner's understanding of the task; (ii) it requires her to be familiar with MACHINE LEARNING techniques and perform several experiments; and (iii) insights gained from the experiments cannot be efficiently transferred to benefit

ISSN: 2319-8753

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 8, August 2014

other communities. We envision an automated data processing framework that takes the domain data and the end goal as input and abstracts the model selection and adaptation task for the practitioner. This would replace till then subjective and manual process with a principled approach that generalizes and scales to large, varied datasets. The design of this framework is a sub-area within MACHINE LEARNING that can leverage exciting new advances in different Learning techniques like, Active, Structure and Computational Learning Theory.

We propose to invite MACHINE LEARNING researchers and practitioners to brainstorm on various ideas that can realize the envisioned Meta Machine Learning framework. One idea that can be explore is the primary mapping between large-scale data streams and learning algorithms. Different learning methods are known to be well behave in terms of on condition that expected outcomes for a given input. A discriminative classifier such as SVM can classify a linearly separable dataset accurately and a time-series sequence can be accurately modeled by a Markov chain as HMM. These thumb rules about model medley rely on the bias property of the data. We think that the relationship between bias of the data and models is relatively unexplored, notwithstanding some functional results in Computational Learning Theory. Another possible method that can be discussed at the workshop is the study of cross-disciplinary practices of exploring data-model relationships. Same problems have been studied in life sciences where disease patterns are modeled through a set of observations in the form of sensor readings. What associations exist between diseases and annotations and how does one automatically infer this relationship given massive amount of data? Economists also have their individual model indicators in the form of inflation rate and stock prices as variables to define the health of economy. The bond between thousands of variables and the practice of choosing a subset of these for consideration is an art. It needs to be reduced to a scalable scientific technique thereby benefitting many other disciplines as well.

III.GOAL OF MACHINE LEARNING ALGORITHMS ON BIG DATA

- Data Categorization: How does one spot different classes of data that fit a particular learning algorithm?
- Data Representation: How does one combine domain-independent data representation (i.e. features) with domain-specific representations to provide efficient learning results?
- Model Refinement: How does one recuperate learning models for each data category?
- Model Formalism: How do we sanctify learning in a framework that can be rapidly deployed across diverse disciplines dealing with large-scale data? Can we treat models as objects?
- Scripting Languages for MACHINE LEARNING: Does it help to design simple scripting languages for computations on learning models in MACHINE LEARNING framework.

IV.APPLICATIONS IN ECONOMICS, ENERGY AND HEALTHCARE

E-commerce web analytics has matured in recent years but is still restrained within individual E- commerce sites. The difficulty with scaling the techniques across multiple sites has been a challenge due to non-uniformity of the features which form the basis for learning website visitor behaviors. The proposed MACHINE LEARNING techniques to capably learn user behaviors would enable scalability across multiple sites.

Machine Learning has become common in healthcare applications but a number of challenges are yet to be addressed. For example, finding suitable classifiers is pivotal to fraud prevention and formulating statistical models for disease prophecy is a barrier to accurate diagnosis. MACHINE LEARNING techniques would enable encoding prior knowledge so that given a particular healthcare data; the system would summon the optimal learning routine.

The field of economics has been a complex domain for predictive analytics because:

ISSN: 2319-8753

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 8, August 2014

(1) There is a scarcity of economic signals that can be captured as features on which learning algorithms can be trained. (2) Statistical models are too narrow to capture macroeconomic data sets. MACHINE LEARNING would enable to overcome these challenges and fit the right model for a given application.

Obtainable techniques for smart energy are faced with optimization problems such as what signal should be capture from the energy usage data to design the best demand-response system. The proposed MACHINE LEARNING framework will address such questions in a principled manner.

V.CONCLUSION

As the complexity of enterprise system is increasing, the need for monitoring and analyzing such systems also grows. Using meta algorithm of machine learning for big data a sophisticated monitor tools has been developed. For example, based on instrumentation and specific APIs, it is now possible to monitor single method invocations and trace individual transactions across geographically distributed systems. Big data enables more precise forms of analysis and prediction. To maximize the benefit of data monitoring the data has to be stored for an extended period of time for ulterior analysis. This new wave of big data analytics imposes new challenges for the application performance monitoring systems. The monitoring data has to be stored in a system that can uphold the high data rates and at the same time enable an up-to-date view.

REFERENCES

1. B. Park and H. Kargupta, Distributed data mining: Algorithms, systems, and applications, Distributed data mining handbook.
2. Richard L. Villars, Carl W. Olofson, Mathew Eastwood, june 2011 title of paper is Big Data: What it is and why you should care. Steve Oberlin title of paper is Machine learning, cognition & big data.
3. Alek kolcz, title of paper is large scale machine learning at twitter.
4. Kristin P. Bennett, Emilio Parrado-Hernandez title of paper is The Interplay of Optimization and Machine Learning Research.
5. Tilmann Rabl, Mohammad sadogi, Hans Arno title of paper is Challenges for enterprise application performance management. Schuller, B., Villar, R.J., Rigoll, G. and Lang, M., title of paper Meta-Classifiers in Acoustic and Linguistic Feature Fusion-Based Affect Recognition, In Proceedings of Acoustics, Speech, and Signal Processing, 2005. (ICASSP '05)
6. C.C. Chibelushi, J.S.D. Mason, and F. Deravi, Feature-level data fusion for bimodal person recognition, In Proceedings of 6th International Conference on Image Processing and its Applications, 1997
7. Blum, A., Mitchell, T. combining labeled and unlabeled data with co-training. In Proceedings of the Workshop on Computational Learning Theory, 1998
8. Leo Breiman, Bagging predictors In proceedings of Machine Learning, 1996
9. Yoav Freund and Robert E. Schapire, A decision-theoretic generalization of on-line learning and an application to boosting, In Proceedings of the Second European Conference on Computational Learning Theory, EuroCOLT '95Polikar, R., Ensemble based systems in decision making, In proceedings of Circuits and Systems Magazine, 2006
10. Robert E. Schapire, The Strength of Weak Learnability, In proceedings of Machine Learning, 1990
11. Gates. Programming Pig. O'Reilly, 2011.
12. Gates, O. Natkovich, S. Chopra, P. Kamath, S. Narayanamurthy, C. Olston, B. Reed, S. Srinivasan,and U. Srivastava. Building a high-level dataflow system on top of Map Reduce: The Pig experience. VLDB, 2009.