International Journal of Engineering & Technology, 7 (1.3) (2018) 48-51

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Technical challenges and perspectives in batch and stream big data machine learning

KVSN Rama Rao[1]*, Sivakannan S[2], M.A.Prasad[3], R.Agilesh Saravanan[4]

[1] Dept.of.CSE, Koneru Lakshmaiah Education Foundation, Guntur, Andhra Pradesh, India
[2,4] Dept.of.ECE, Koneru Lakshmaiah Education Foundation, Guntur, Andhra Pradesh, India
[3]Dept.of Comp. Science, Dr.L.Bullayya PG College, Visakhapatnam, Andhra Pradesh, India
**Corresponding author E-mail: kvsnramarao@yahoo.co.in*

Abstract

Machine Learning is playing a predominant role across various domains. However traditional Machine Learning algorithms are becoming unsuitable for majority of applications as the data is acquiring new characteristics. Sensors, devices, servers, Internet, Social Networking, Smart phones and Internet of Things are contributing the major sources of data. Hence there is a paradigm shift in the Machine learning with the advent of Big Data. Research works are in evolution to deal with Big Data Batch and stream real time data. In this paper, we highlighted several research works that contributed towards Big Data Machine Learning.

Keywords: *Big Data, Knowledge Discovery, Machine learning, Batch, Stream*

1. Introduction

Machine Learning (ML) is a Science, which has its origins from Pattern Recognition and Artificial Intelligence. ML is playing vital role in domains such as Data Mining, Expert Systems, and Natural Language Processing. ML primarily encompasses building of algorithms, which involves the tasks of Learning and Prediction. In a way, ML gives the strength to the machines without explicitly programming. For certain tasks such as Spam filtering and search engines, designing specific algorithms may not be feasible, as the requirements will be changing dynamically from time to time. In such cases Machine learning is really a boon. But, in the recent past, data is acquiring diverse characteristics. The size of data has boosted from terabyte to petabyte range. Most of the data is unstructured since it is coming from diversified sources. For example in case of a Network system, the data will be from variety sources such as Firewalls, Intrusion Detection Systems, Clients, Servers, Antivirus software, CC Cameras etc. Each source will have its own format. Dealing these unstructured datasets is a challenge .On the other hand, mining real time data is critical, as it changes with rapid speed. This gave rise to the evolution of Big Data Mining.

2. Bigdata mining

Latest technological developments have resulted in the generation of huge data from sensors, devices, servers etc. Internet, Social Networking, Smart phones and Internet of Things are contributing the major sources of data. Internet of Things (IoT) has revolutionized our life by connecting several devices. Since the applications of IoT spans across different applications such as traffic, weather, home devices control etc. will generate vast amount of data. There are certain challenges in utilizing such enormous amount of data in terms of models and algorithms design.The data generated by these autonomous sources is heterogeneous. This gave rise to the birth of the term Big Data. Doug Laney[4] has defined the characteristics of Big Data in 3V's.

Knowledge Discovery from such Big data with the said characteristics (3V's) is cumbersome and certainly need a different approach/principles. Begoli, Edmon, and James Horey [5] have presented three design principles which assist in effective Knowledge discovery from Big Data.

The Principles are:

a)The process of Machine learning and Data Analysis in case of Big Data should be supported by variety of statistical techniques and a large set of data mining techniques. For Instance supporting R, SAS, Python, MADLib and Hadoop mahout Library.

b) Data Analysis Pipeline for Data Preparation, Processing (both structured and unstructured). This principle emphasizes that a single style database will not cater all the needs. So while dealing with Big Data, tools such as Hadoop, Hive, HBase, Cassandra,Neo4j, PostGIS, GeoTools can be used.

c) By using popular and open standards, creating interfaces and exposing the results through the API.

Wu, Xindong, et al. [6] has presented HACE theorem that proposes Big data processing model from data mining viewpoint. HACE is abbreviated as Heterogeneous (H), Autonomous (A), Complex(C) and Evolving (E) relationships among data.

Heterogeneous Data: This element emphasizes that data is represented in heterogeneous and varied dimensions. In case of medical field, for X-Rays, CT-Scan images and videos are examined while for DNA, genetic information is used. This clearly shows that, different representations are possible for a single individual. Aggregation of the sources of data will be challenge as different practitioners have their own way of representing patient.

Autonomous Sources: Data is being generated by autonomous sources i.e. each data source is decentralized and can generate information.

Complex and Evolving: As the volume of data increases, complexity also increases proportionately. This will make difficult to identify evolving relationship within data.

Further Wu, Xindong, et al[6] proposed a 3 tier Big data Processing framework to provide solution to the above mentioned HACE theorem characteristics. Such kind of framework can address the Big Data Mining issue.

3. Machine learning and bigdata

Big Data Mining is facing the problem of many features making any machine learning algorithm inefficient. Reducing the number of dimensions or features will certainly improve the performance of any machine learning algorithm. Wu, Xindong, et al. [7] presented Online feature selection framework and OSFS algorithms.

Two algorithms are proposed to implement the framework.

a) OSFS algorithm: It is a two phase algorithm performing online relevancy and redundancy analysis.

Phase 1: In relevance analysis phase, strong and weak relevant features are identified and added to set. The algorithm assesses the relevance of new feature with an attribute. Accordingly, decides to discard or add the feature.

Phase 2: Then phase 2 commences which dynamically identifies and removes the redundant features.

The two phases run, till pre-defined accuracy is satisfied or specified iterations reached or no features.

But, redundancy analysis phase is time consuming in OSFS algorithm. Hence they proposed fast OSFS algorithm.

b) Fast OSFS algorithm: This algorithm handles redundancy analysis in two parts. Part one eliminates a new relevant feature but redundant from inclusion in set. If the feature is not eliminated in first part, second part performs a validation check to ensure that any redundancy is introduced due to the addition of new feature in the set. So fast OSFS reduces computation cost by performing tests only on some subsets instead of all in the set.

Authors performed experiments on Mars Impact crater data set and compared the results with state of art traditional algorithms. The results demonstrated that these two algorithms are selecting less features when compared to other methods with good prediction accuracy.

Hoi, S. C.et.al.[8] has presented their work on Online feature selection for Mining Big data. Authors have proposed algorithms to address online feature selection by applying on large data sets. Authors modified the perceptron algorithm by applying truncation and evaluated the performance of algorithms on Big data. Each dataset is considered with at least one lakh instances (KDDCUP08 data set with 102294 number and 117 dimensions. Codrna dataset with 271617 number 8 dimensions. Covtype Dataset with 581012 number and 54 dimensions).The experiments were performed on a normal PC with a dual core Processor and is implemented in Matlab. The OFS algorithm is evaluated with standard algorithms. OFS algorithm i.e. the projected algorithm out rates the random and truncated algorithms. But the time taken in learning the features on large data sets is equally efficient.

Kraska, T et al. [9] have presented their idea ML-base, a distributed ML system. ML-base addresses a broad range of users and large data sets.ML-base provides functionality for several ML tasks such as classification, regression, feature selection and dimension reduction. ML base has a declarative language which can be applied across use cases. One use case is to predict a Neuro illness and observe whether the patient is exhibiting delayed disease progression. Another use case is to predict the song they are listening. This prediction will be based on listening history and training set. This comes under the example of collaborative filtering, where a matrix with columns as users and songs as rows is created. Infer from the entries. Authors presented their architecture which consists master and slave nodes. User submits his request using ML descriptive language which is parsed into a learning plan. The learning plan consists of several algorithms and techniques. Then the optimizer comes into picture and generates and optimized learning plan. Subsequently, this will be converted to a physical learning plan by clearly specifying features, Map reduce operations, datasets etc. to be used. Now the master distributes these tasks to nodes. A learned model is returned as a result that the user may use for predictions. The model is further improved via additional exploration which will further improve the learning plan. In summary, this model first transforms the ML task into logical plan. Then optimize it and generate physical plan. This algorithm involves two methods. One is Gradient Computation which considers the data and parameters of the model and calculate the gradient. Second method is Update function which maps the current parameters to new parameters using the computed gradient. Authors observed that Gradient descent algorithms are robust and possess statistical freedom. Finally authors concluded that, MLbase is fully distributed and has runtime ability to exploit ML algorithms. Authors reported their primary results depicting the strength of Optimizer in this paper.

Lin, J., &Kolcz, A [10] has presented a case study on how twitter has integrated machine learning tools to Hadoop/Pig centric platform. Huge volume and variety of data will be stored in HDFS. Java code is written to interact with HDFS. But at Twitter, most of the code is written using a high level data flow language called Pig [11][12]. Author’s contribution lies in how to integrate machine learning capabilities into Pig. To achieve this, a framework with two components such as Java Library and light weight wrappers are developed. Core java library contains classifiers, trainers and also interface to connect to third party packages. For model training, java feature vectors are exposed as maps in Pig.

For large scale machine learning, stochastic gradient descent (SGD) and ensemble methods are widely used. SGD is a linear model that outputs an estimate of posterior probabilities. Authors applied their ML model on sentiment analysis. To generate training data for classifying as positive or negative, emotions are used. Emotions with positive indications are treated as positive and remaining as negative. They collected one million tweets with emotions as test set. For training, one million, 10million and 100 million example tweets are collected. Now emotions are removed from both training and test sets. Logistic regression classifier is used with SGD. The feature extractor is hashed byte 4-grams where a sliding window moves over the array and generates a hash, which becomes the value of feature id. For training,1,10,100 Million datasets, a single reducer is used and for 10,100 million datasets with multiple reducers (ensemble). For ensemble model, the probabilities of all models are taken into consideration. The accuracy is better upto 10 million examples and is slightly better in case of 10 to 100 million examples. Also reported that, ensembles performed well rather than a single classifier. Ensembles take shorter time to train. This paper presented a solution to classification, however a similar kind of method applies to clustering which can be taken up as a future work. Sutharan, S.[13] has emphasized that many domains suffer from Big Data issue such as business, geo space. Networking also joined the list. Detecting and Predicting a Network Intrusion is a time sensitive application which require high end tools and techniques. The three parameters pose a challenge to networking domain. To tackle, Network topology must be designed cost effectively. A better solution is to integrate cloud and HDFS. Accordingly, author proposed a model consisting of four units: User interaction, Network traffic recording, HDFS and Cloud. Authors adopted Cross Domain, Representation Learning (CDRL) technique by Tu and Sun [14] for classification. But the implementation of this will raise challenges such as constructing geometric representation, extracting suitable features. To address this Unit Circle Algorithm (UCA) is used, where traffic data is represented by unit circles and assigns many related records to fewer circles. NSL-KDD dataset was used for experiments. The author concluded by leaving few future works such as validation of learned knowledge, unnecessary repetition of learning process for new data

Sun, Y., & Han, J [16] has presented their methodology that can effectively mine knowledge from heterogeneous information networks. Few examples can be social networks, World Wide Web etc. For instance in case of DBLP, different kinds of knowledge can be derived such as clustering, ranking, classification, topic analysis etc. These functions facilitate generation of new knowledge. The interesting fact is most of the real world networks are heterogeneous where nodes and relations are of different types. Treating all nodes as same type or every node as same type will lose valuable information. For example, treating all nodes such as patients, doctors, tests, medicines etc. as same type is incorrect. Similarly, we should consider all patients are of same kind and doctors are of different kind. Similarly in Facebook, it consists of persons as well as objects of other types, such as photos, posts, companies, and movies. This semi structure heterogeneous network will provide us with essential information. Authors have summarized several principles that will be helpful for heterogeneous information analysis. Metapath based search can be used for searching and mining. It is under the assumption that objects are connected via different paths. In case of paper publication, example of Meta paths can be author-paper-author &author-paper-venue-paper-author" path. User guidance will lead to strength aware mining. Since there will be different meta paths which represent different relations with different semantic meanings, user guidance training examples can better guide for clustering. The weighted meta paths can be learned to have better consistency.

Few advanced topics in mining like Role discovery, Trustworthiness Analysis, Text Mining and Information networks Integration and OLAP are discussed by the author in case of heterogeneous networks.

The authors concluded by giving several research directions:

- entity extraction, data cleaning, detection of hidden semantic relationships, and trustworthiness analysis should be integrated with the network construction and mining processes to progressively and mutually enhance the quality of construction and mining of information networks.
- Information spreading models in the heterogeneous networks to be studied.
- A user may be interested only on a certain part of information. It is better to mine that part only, instead of entire network. For example mine only suspected networks and their links.

Lin, J., & Ryaboy, D [17] has presented their infrastructure and development of capabilities in Mining Big Data. Authors had good real time working experience at twitter. Big data analysis Life Cycle commences with understanding of data. From the daily activities, lots of data will be generated but it will be more confused. For example, users login and logouts, what features of a product do users use, activity profiles changing over time etc. Before beginning analysis, analyst should understand what data is available and how they are organized. To understand this, we need to have better insights on service oriented architectures and other issues.

The operation logs of a complex service operation are adhoc. A well-structured log should be designed to capture information such as request initiator, time of request, result of operation, how long, any warnings or exceptions etc., These logs provide valuable information to data Mining.

Another biggest challenge is handling impedance mismatch between different systems and frameworks. Each framework will be effective for a certain problem. Choosing right tool for the right job makes sense. Each framework constructs models and sets the direction of thinking. For example, map Reduce forces the developer to think in terms of Maps and Reduces. Further there will be threading between data flows. For the complete data, links has to be drilled downwards. Orchestrating this entire process is complex because data sources may be heterogeneous originating from different frameworks. These issues affect data mining.

The authors concluded by throwing several research issues to be explored.

- Visualization for generating insights. Tool like d3.js is limited to engineer laptop and is constrained by browser ability to deal with large datasets.
- Real-time interactions with large datasets. Hadoop based is not able to provide shorter data mining cycles.

4. Knowledge discovery from streams

Gama, J.[18] in his book has given a detailed discussion on Data streams and techniques. Data streams can be seen as stochastic processes in which events occur continuously and independent from each other. Some characteristics of data streams include transient streams, continuous queries, sequential access, unpredictable data characteristics and arrival patterns. Author has presented a good discussion on change detection, streaming algorithms, clustering from data streams, Decision trees and novelty detection. On Time series also author presented nice discussion.

Bifet, A[19] has discussed about current and future trends of mining evolving data streams. Real time data stream analytics is the need of the hour with applications ranging from sensors, traffic management, manufacturing, twitter, etc. Data streams arrive at very high speed and the algorithms processing must work under rigid space and time constraints. In other words, data stream mining concentrates on three dimensions such as accuracy, memory and time to predict. The challenge is all three dimensions are independent. If we look to improve one dimension, other dimension may suffer. Estimating the combined cost of performing learning and prediction in terms of time and memory has come into existence. Cost per hour usage in terms of RAM hours is the measure. GB of RAM deployed for an hour is one RAM hour. It is used as a measure to know the resources used by the streaming algorithm. Big data analytics can be performed using several open source tools such as Hadoop, Pig, Cassandra, Storm, HBase etc. For Big data Mining, mahout, MOA, R, Wabbit, Pegasus, Graphlab can be used. We will explore about MOA further.

Bifet, A.[20] has described MOA (Massive Online Analysis) which is a software environment for implementing algorithms to learn from evolving data streams. They introduced Green Computing which is a study and practice of using computing resources efficiently.MOA can be integrated with WEKA. The Life Cycle of a data stream classification comprises the following:

- The algorithm is given the next available stream.
- The algorithm processes and updates without exceeding its memory bounds. It also completes quickly.
- Now it is able to predict unseen examples on request.

MOA permits evaluation of classification on large streams on over ten million examples. MOA is written in java to have the advantage of portability.

De Francisci Morales, G.[21] has presented SAMOA (Scalable Advanced Massive Online Analysis). Author reiterated that Big data streams are characterized by volume, velocity and sometimes variety. Several areas need to be combined to deal with such complex data.

- Distributed Computing to deal with volume of data
- Open Source tools to deal with variety
- Streaming paradigm to deal with velocity

MOA is a framework for data stream mining, although for single machines. SAMOA is a platform for Big data stream Mining and Machine Learning. SAMOA supports classification and clustering. SAMOA is an Open source.

Amatriain, X[22] has considered Netflix use case and described different kinds of machine learning techniques for mining large streams. Recommender systems are better examples of large scale stream data mining. E –commerce, music, video, are few applications which involve large volume mining and generate data as per user personalization need. One frequently used approach for Recommender system is Collaborative Filtering (CF) algorithms.

The authors concluded by emphasizing that, recommendation problem are far from solved. It is because we need to take into consideration several factors such as context, popularity, interest, evidence, novelty, diversity, or freshness. Supporting all the different contexts in which we want to make recommendations requires a range of algorithms and different kinds of data.

5. Conclusion

Big Data Machine Learning is an upcoming research field. Volume, Variety and Velocity data characteristics initiate the need for development of new algorithms and models for machine Learning. Several research works have been discussed in this paper which gave directions for the development of Big data machine Learning Algorithms. But No single method or algorithm can function efficiently due to the data characteristics. Batch Processing Machine Learning algorithms were addressed to certain extent but real time stream Mining is still an open area.

References

[1] http://www.cs.cmu.edu/~tom/

[2] I. Witten, E. Frank, and M. Hall. Data Mining:Practical Machine Learning Tools and Techniques.Morgan Kaufmann, San Mateo, CA, 3rd edition, 2011.

[3] Domingos, Pedro. "A few useful things to know about Machine Learning."Communications of the ACM 55.10 (2012): 78-87.

[4] D. Laney. 3-D Data Management: Controlling Data Volume, Velocity and Variety. META Group Research Note, February 6, 2001

[5] Begoli, Edmon, and James Horey. "Design principles for effective knowledge discovery from big data." Software Architecture (WICSA) and European Conference on Software Architecture (ECSA), 2012 Joint Working IEEE/IFIP Conference on. IEEE, 2012.

[6] Wu, Xindong, et al. "Data mining with big data." Knowledge and Data Engineering, IEEE Transactions on 26.1 (2014): 97-107.

[7] Wu, Xindong, et al. "Online feature selection with streaming features."Pattern Analysis and Machine Intelligence, IEEE Transactions on 35(5) (2013): 1178-1192.

[8] Hoi, S. C., Wang, J., Zhao, P., & Jin, R. (2012, August). Online feature selection for mining big data. In Proceedings of the 1st international workshop on big data, streams and heterogeneous source mining: Algorithms, systems, programming models and applications (pp. 93-100). ACM.

[9] Kraska, T., Talwalkar, A., Duchi, J. C., Griffith, R., Franklin, M. J., & Jordan, M. I. (2013). MLbase: A Distributed Machine-learning System. In CIDR.

[10] Lin, J., &Kolcz, A. (2012, May). Large-scale machine learning at twitter. InProceedings of the 2012 ACM SIGMOD International Conference on Management of Data (pp. 793-804). ACM.

[11] C. Olston, B. Reed, U. Srivastava, R. Kumar, and A. Tomkins. Pig Latin: A not-so-foreign language for data processing. SIGMOD, 2008.

[12] A. 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 MapReduce: The Pig experience.VLDB, 2009.

[13] Suthaharan, S. (2014). Big data classification: Problems and challenges in network intrusion prediction with machine learning. ACM SIGMETRICS Performance Evaluation Review, 41(4), 70-73.

[14] Tu, W., & Sun, S. (2012, August). Cross-domain representation-learning framework with combination of class-separate and domain-merge objectives. In Proceedings of the 1st International Workshop on Cross Domain Knowledge Discovery in Web and Social Network Mining (pp. 18-25). ACM.

[15] Kang, U., &Faloutsos, C. (2013). Big graph mining: algorithms and discoveries. ACM SIGKDD Explorations Newsletter, 14(2), 29-36.

[16] Sun, Y.,& Han, J.(2013). Mining heterogeneous information networks: a structural analysis approach. ACM SIGKDD Explorations Newsletter, 14(2), 20-28.

[17] Lin, J., & Ryaboy, D. (2013). Scaling big data mining infrastructure: the twitter experience. ACM SIGKDD Explorations Newsletter, 14(2), 6-19.

[18] Gama, J. (2010). Knowledge discovery from data streams. CRC Press.

[19] Bifet, A. (2013). Mining big data in real time. Informatica, 37(1).

[20] Bifet, A., Holmes, G., Kirkby, R., &Pfahringer, B. (2010). Moa: Massive online analysis. The Journal of Machine Learning Research, 11, 1601-1604.

[21] De Francisci Morales, G. (2013, May). SAMOA: A platform for mining big data streams. In Proceedings of the 22nd international conference on World Wide Web companion (pp. 777-778). International World Wide Web Conferences Steering Committee.

[22] Amatriain, X. (2013). Mining large streams of user data for personalized recommendations. ACM SIGKDD Explorations Newsletter, 14(2), 37-48.