PROPOSAL OF CONCEPTUAL MODEL FOR SUCCESSFUL IMPLETION OF MACHINE LEARNING WITH BIG DATA

Ganga Patur
Research Scholar
Shri JJT University
Rajasthan
ganga06patur@gmail.com

Dr.K.E.Balachandrudu
Professor
Arjun College of
Technology & Sciences
Hayathnagar

Dr. S.K.Yadav
Professor
Shri JJT University
Rajasthan

Abstract:

The Big Data revolution promises to transform how we live, work, and think by enabling process optimization, empowering insight discovery and improving decision making. The realization of this grand potential relies on the ability to extract value from such massive data through data analytics; machine learning is at its core because of its ability to learn from data and provide data driven insights, decisions, and predictions. However, traditional machine learning approaches were developed in a different era, and thus are based upon multiple assumptions, such as the data set fitting entirely into memory, what unfortunately no longer holds true in this new context. These broken assumptions, together with the Big Data characteristics, are creating obstacles for the traditional techniques. Consequently, this paper compiles, summarizes, and organizes machine learning challenges with Big Data. Moreover, emerging machine learning approaches and techniques are discussed in terms of how they are capable of handling the various challenges with the ultimate objective of helping practitioners select appropriate solutions for their use cases. Finally, a matrix relating the challenges and approaches is presented. Through this process, this paper provides a perspective on the domain, identifies research gaps and opportunities, and provides a strong foundation and encouragement for further research in the field of machine learning with Big Data.

Key words:

machine learning, traditional techniques, Big Data, further research, organizes, challenges.

Introduction:

today, the amount of data is exploding at an unprecedented rate as a result of developments in Web technologies, social media, and mobile and sensing devices. These Big Data possess tremendous potential in terms of business value in a variety of fields such as health care, biology, transportation, online advertising, energy management, and financial services. However, traditional approaches are struggling when faced with these massive data.

The concept of Big Data is defined by Gartner as high volume, high velocity, and/or high variety data that require new processing paradigms to enable insight discovery, improved decision making, and process optimization. According to this definition, Big Data are not characterized by specific size metrics, but rather by the fact that traditional approaches are struggling to process them due to their size, velocity or variety. The potential of Big Data is highlighted by their

definition; however, realization of this potential depends on improving traditional approaches or developing new ones capable of handling such data.

Because of their potential, Big Data have been referred to as a revolution that will transform how we live, work, and think. The main purpose of this revolution is to make use of large amounts of data to enable knowledge discovery and better decision making.

Machine Learning Challenges Originating From Big Data Definition :

Big Data are often described by its dimensions, which are referred to as its Vs. Earlier definitions of Big Data focussed on three Vs (volume, velocity, and variety); however, a more commonly accepted definition now relies upon the following four Vs: volume, velocity, variety, and veracity. It is important to note that other Vs can also be found in the literature. For example, value is often added as a 5th V. However, value is defined as the desired outcome of Big Data processing and not as defining characteristics of Big Data itself. For this reason, this paper considers only the four dimensions that characterize Big Data Fig. 1 illustrates the dimensions of Big Data along with their associated challenges as further discussed in the following sub-sections.

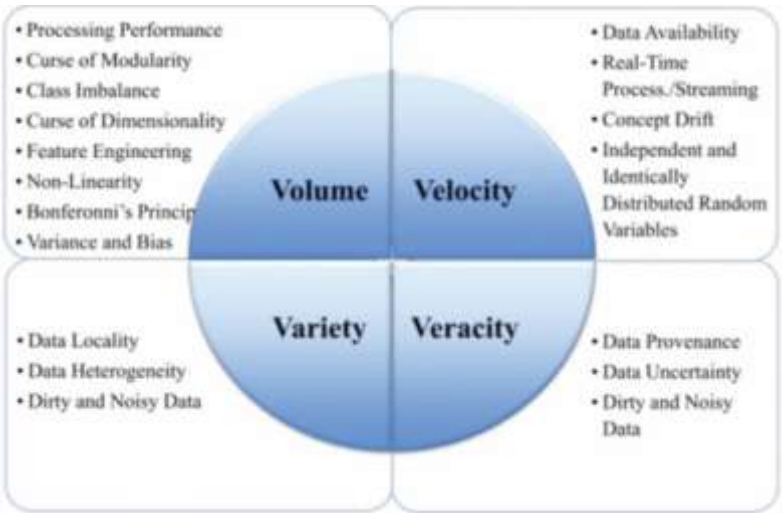

Fig. 1.:Big Data characteristics with associated challenges.

Approaches:

In response to the presented challenges, various approaches have been developed. Although designing entirely new algorithms would appear to be a possible solution, researchers have mostly preferred other methods. Many approaches have been suggested and surveys have been published on specific

categories of solutions; examples include surveys on platforms for Big Data analytics and review of data mining with Big Data. This paper reviews and organizes various proposed machine learning approaches and discusses how they address the identified challenges. The big picture of approach-challenge correlations is presented in Table 1; it includes a list of approaches along with the challenges that each best addresses. Symbol ‘’ indicate high degree of remedy while ‘*’ represents partial resolution.

APPROACHES			CHALLENGES: VOLUME: Processing Performance	VOLUME: Curse of Modularity	VOLUME: Class Imbalance	VOLUME: Curse of Dimensionality	VOLUME: Feature Engineering	VOLUME: Non-linearity	VOLUME: Bonferonni's Principle	VOLUME: Variance and Bias	VARIETY: Data locality	VARIETY: Data Heterogeneity	VARIETY: Dirty and noisy Data	VELOCITY: Data availability	VELOCITY: Real-time Processing/Streaming	VELOCITY: Concept drift	VELOCITY: I.i.d	VERACITY: Data Provenance	VERACITY: Data Uncertainty	VERACITY: Dirty and Noisy Data
MANIPULATIONS	Data Manipulations	Dimensionality Reduction	✓			✓														
		Instance Selection	✓	✓																
		Data Cleaning											✓							✓
	Processing Manipulations	Vertical Scaling	✓															*		
		Hori-zontal Scaling: Batch-oriented	✓	✓		*					✓							*		
		Hori-zontal Scaling: Stream-oriented	✓	✓										✓	✓			*		
	Algorithm Manipulations	Algorithm Modifications	✓	*		*					✓				✓					
		Algorithm Mod. with new Paradigm	✓	*		*					✓				✓					
LEARNING PARADIGMS		Deep Learning					✓	✓				✓	*						*	*
		Online Learning	✓	✓	*						✓		*	✓	✓	*	✓			*
		Local Learning	✓	✓	✓					✓	✓									
		Transfer Learning			✓							✓	*						*	*
		Lifelong Learning	✓		✓							✓	*	✓	✓	*			*	*
		Ensemble Learning	✓	✓												✓				

TABLE 1 :Machine Learning Approaches and the Challenges they Address

As it can be seen from the table, there are two main categories of solutions. The first category relies on data, processing, and algorithm manipulations to handle Big Data. The second category involves the creation and adaptation of different machine learning paradigms and the modification of existing algorithms for these paradigms.

The following sub-sections introduce techniques and methodologies being developed and used to handle the challenges associated with machine learning with Big Data. First, manipulation techniques used in conjunction with existing

algorithms are presented. Second, various machine learning paradigms that are especially well suited to handle Big Data challenges are discussed.

A. Manipulations for Big Data:

Data analytics using machine learning relies on an established suite of events, also known as the *data analytics pipeline*. The approaches presented in this section discuss possible manipulations in various steps of the existing pipeline. The purpose of these modifications is to respond to the challenges of machine learning with Big Data. Fig. 3 shows a representation of the pipeline based on the work of Labrinidis and Jagadish, along with the three types of manipulations to be discussed in this section: data manipulations, processing manipulations, and algorithm manipulations. These three categories, along with their corresponding sub-categories and sample solutions, are presented in Fig. 4. The examples included are only representatives and in no way provide a comprehensive list of solutions.

Fig. 2.Data Analytics Pipeline.

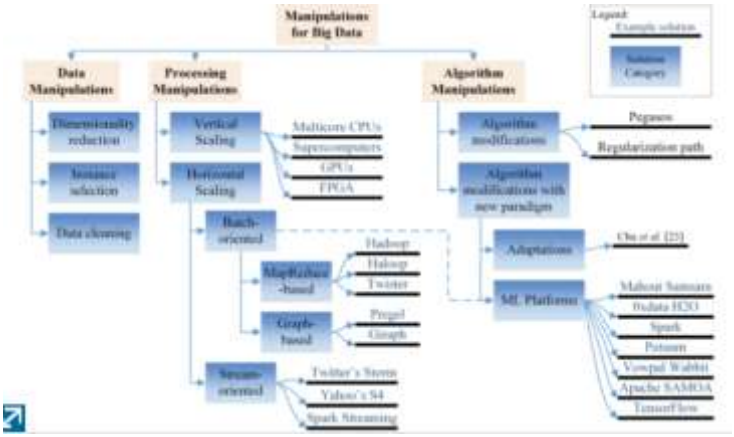

Fig. 3.Manipulations for Big Data.

1. **Data manipulatons:**

One of the first manipulations to be attempted in an effort to adapt Big Data for machine learning is to try to modify the data in order to mimic non-Big Data. This modification takes place in the data pre-processing stage of the pipeline, as illustrated in Fig. 3.

Two of the most important data-related aspects affecting machine learning performance are high dimensionality (wide datasets) and large number of samples (high datasets). Therefore, two intuitive data manipulations for learning with Big Data are dimensionality reduction and instance selection as shown in Fig. 4. The term ***data reduction*** sometimes refers to both these manipulations, but occasionally specifically denotes instance selection. Additionally, data clearing is another important aspect of data manipulation.

Dimensionality reduction aims to map high dimensionality space onto lower-dimensionality one without significant loss of information. A variety of means exists to reduce dimensions in the context of Big Data. One popular, but very old technique (it originates from 1901) is *principal component analysis* (PCA). PCA belongs to the family of linear mapping techniques: orthogonal transformations are applied to transform a set of possibly correlated variables into a set of linearly uncorrelated variables, called principal components. The first principal component accounts for the largest proportion of the variability in the data, the second one has the next highest variance and is orthogonal to the first, and so on. Thus, choosing only the first p principal components can reduce dimensionality.

Instance selection refers to techniques for selecting a data subset that resembles and represents the whole dataset. Whereas dimensionality reduction deals with wide datasets, data reduction, more specifically instance selection, aims to reduce a dataset's height. The subset is consequently used to make inferences about the whole dataset.

Although instance selection reduces dataset size thus improves processing performance and eases the curse of modularity, a number of questions arise:

- How big should the sample be? The sample size should balance accuracy and computing time.

- What sampling approach should be used? The choice of approach has a major impact on how well the subset represents the whole.

- How good will the model be? Instance selection introduces sampling error due to the differences between the sample and the whole dataset.

These issues, although well researched in learning with small datasets, are enlarged in the Big Data context because data size makes it more difficult to evaluate different properties or models.

Moreover, as already mentioned, in the Big Data context, challenges of class imbalance, noise, variance, and bias are more common and more difficult. In turn, this makes it more challenging to select a subset that will adequately represent the whole set. For example, with a large class imbalance, the selection approach must ensure that instances from all classes are selected. On the other hand, an appropriate instance selection can remedy class imbalance.

Data cleaning is another type of data manipulations; it refers to pre-processing such as noise and outlier removal. Thus, it tackles the challenges of dirty and noisy data. In this area, there is no significant development with respect to Big Data. Noise removal has been an especially active research topic in the audio, image, and video domains.

2) Processing Manipulations

To improve machine learning performance with Big Data, processing manipulations focus on modifying how data are processed and stored. Here, the term *storage* refers not only to physical storage on a permanent medium, but also to how data are represented in memory. As illustrated in Fig. 3, processing manipulations can happen during three phases of the data analytics pipeline: data transformation, data storage, and data analysis.

In these stages, independent of the category of manipulations, processes can be embedded to capture data provenance and therefore remedy provenance challenge; Many learning algorithms, such as brute-force search and genetic algorithms, are trivially parallel, and therefore parallelization can provide massive performance improvements. Consequently, researchers have developed techniques and tools to parallelize machine learning.

This paradigm has been developed in two streams: batch- and stream-oriented systems.

Batch-oriented systems process a large amount of data at once, have access to most of the data, and typically are more concerned with throughput than with latency. MapReduce-based solutions address the curse of modularity as they typically do not require the complete dataset to be held in memory. Moreover, data locality is also resolved as those solutions support work with data residing on different physical location. Such solutions facilitate work with high dimensional data, but they do not resolve the breakdown of the similarity-based reasoning, thus they provide partial resolution for the curse of dimensionality.

Stream-oriented systems operate on one data element or a small set of recent data in real-time or near real-time. Both Storm and S4 express computations using a graph topology, and their runtime engines handle parallelization, message passing, and fault tolerance. In contrast to Storm and S4, which perform one-by-one processing, Spark Streaming divides data into micro-batches and carries out computation on the micro-batches.

Research in ML with Big Data has focussed mainly on the horizontal scaling paradigm: MapReduce-based solutions, graph-based solutions, and streaming. As discussed earlier, each category addresses specific problems and encounters difficulties with others. All solutions from the processing manipulation category primarily focus on improving performance (throughput or latency) and do not remedy a number of other challenges as illustrated in Table 1. The combination of processing manipulations with algorithms and new learning paradigms provides research opportunities to undertake the remaining Big Data challenges.

3) Algorithm Manipulations

Algorithm manipulations include approaches that modify existing algorithms, with or without applying new paradigms. Since the very beginning of machine learning, researchers have been trying to improve existing algorithms and to reduce their time and/or space complexity. With Big Data, these efforts have intensified because it has become more important to handle large datasets.

Algorithm modifications have focussed on modifying algorithms to improve their performance. For example, the following approaches have been developed for specific machine learning algorithms to address volume challenges:

- Pegasos provides an optimized version of the support vector machine (SVM) algorithm for large-scale text processing. Its runtime does not depend directly on training set size, and hence Pegasos is especially suitable for large datasets.

- Regularization paths for linear models supports linear regression, two-class logistic regression, and multinomial regression problems. This approach enables processing of large datasets and efficiently handles sparse features.

Solutions from this category deal with processing performance and real-time processing. As they are typically distributed computing solutions, they also address the data locality challenge. Moreover, the curses of dimensionality and modularity are partially remedied as those solutions support work with high dimensional data and may provide smaller memory footprint.

Algorithm modifications with new paradigms category involves modifying ML algorithms to work better with new process manipulations and/or new paradigms. An example from this category would be to modify an algorithm through parallelization and to adapt the algorithm for a new parallel processing paradigm such as MapReduce. Chu *et al.* adapted several algorithms to multicore MapReduce, including naïve Bayes, Gaussian discriminative analysis, k -means, neural networks, support vector machines, and others. All algorithm modification solutions (with or without new paradigms) focus on providing the capability to process large datasets, improving performance, or providing real-time processing capabilities. They also remedy data locality as distributed computation can be performed with data residing on different physical locations.

B. Machine Learning Paradigms for Big Data

A variety of learning paradigms exists in the field of machine learning; however, not all types are relevant to all areas of research. For example, Deng and Li presented a number of paradigms that were applicable to speech recognition. Congruently, the work presented here includes machine learning

paradigms relevant in the Big Data context, along with how they address the identified challenges.

1) Deep Learning

Deep learning is an approach from the representation learning family of machine learning. Representation learning is also often referred to as feature learning . This type of algorithm gets its name from the fact that it uses data representations rather than explicit data features to perform tasks. It transforms data into abstract representations that enable the features to be learnt. In a deep learning architecture, these representations are subsequently used to accomplish the machine learning tasks. Henceforth, because the features are learned directly from the data, there is no need for feature engineering. In the context of Big Data, the ability to avoid feature engineering is regarded as a great advantage due to the challenges associated with this process.

Deep learning uses a hierarchical learning process similar to that of neural networks to extract data representations from data. It makes use of several hidden layers, and as the data pass through each layer, non-linear transformations are applied. These representations constitute high level complex abstractions of the data . Each layer attempts to separate out the factors of variation within the data. Because the output of the last layer is simply a transformation of the original input, it can be used as an input to other machine learning algorithms as well. Deep learning algorithms can capture various levels of abstractions, thus this type of learning is an ideal solution to the problem of image classification and recognition. Fig. 5 provides an abstract view of the deep learning process . Each layer learns a specific feature: edges, corners and contours, and object parts.

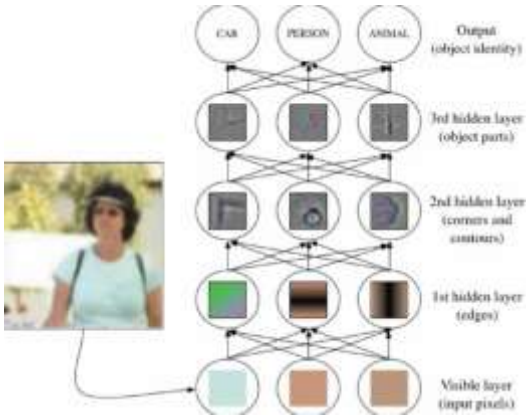

Fig. 5. Deep Learning [101].

They work through backpropagation by attempting to set their target output as their input, thereby auto-encoding themselves. Interestingly, deep learning can be used for both supervised and unsupervised learning. This is possible due to the very nature of the technique; it excels at extracting global relationships and patterns from data because of its reliance upon creating high level abstractions

2) Online Learning

Because it responds well to large-scale processing by nature, online learning is another machine learning paradigm that has been explored to bridge efficiency gaps created by Big Data. Online learning can be seen as an alternative to batch learning, the paradigm typically used in conventional machine learning. As its name implies, batch learning processes data in batches and requires the entire dataset to be available when the model is created . Furthermore, once generated, the model can no longer be modified. This makes it difficult to deal with the dimensions of Big Data for the following reasons:

- Volume: having to process a very large amount of data at one time is not computationally efficient or always feasible.
- Variety: the need to have the entire dataset available at the beginning of the processing limits the use of data from various sources.
- Velocity: the requirement to have access to the entire dataset at the time of processing does not enable real-time analysis or use of data from various sources.
- Veracity: because the model cannot be altered, it is highly susceptible to performance impediments caused by poor data veracity.

Conversely, online learning uses data streams for training, and models can learn one instance at a time.

Furthermore, the descriptor “online” also reflects the fact that this paradigm continuously maintains its model; the model can be modified whenever the algorithm sees fit. Its adaptive nature makes it possible to handle a certain amount of dirty and noisy data, class imbalances, and concept drift. Indeed, Mirza *et al.*

3) Local Learning

First proposed by Bottou and Vapnik in 1992, local learning is a strategy that offers an alternative to typical global learning. Conventionally, ML algorithms make use of global learning through strategies such as generative learning. This approach assumes that based upon the data's underlying distribution, a model can be used to re-generate the input data. It basically attempts to summarize the entire dataset, whereas local learning is concerned only with subsets of interest. Therefore, local learning can be viewed as a semi-parametric approximation of a global model. The stronger but less restrictive assumptions of this hybrid parametric model yield low variance and bias.

Fig. 6 provides an abstract view of the local learning process. The idea behind it is to separate the input space into clusters and then build a separate model for each cluster. This reduces overall cost and complexity. Indeed, it is much more efficient to find a solution for k problems of size m/k than for a single problem of size m. Consequently, such approach could enable processing of datasets that were considered too large for global paradigms. Another way of implementing local learning is to modify the learning algorithms so that only neighbouring samples influence the output variable.

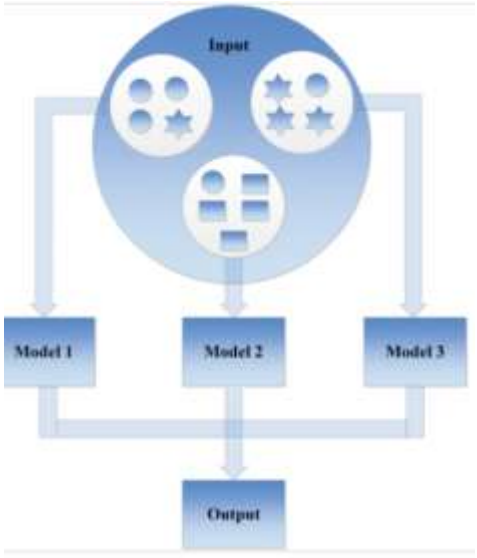

Fig. 6. Local Learning.

Therefore, the challenge of the curse of modularity, class imbalance, variance and bias, and data locality can be alleviated by a local approach. However, matters of dimensionality and velocity, such as concept drift among others, have yet to be addressed. Overall, in the Big Data context, the local approach remains

largely unexplored; studying how this paradigm could better handle velocity and veracity challenges appears to be particularly open.

4) Transfer Learning

Transfer learning is an approach for improving learning in a particular domain, referred to as the *target domain*, by training the model with other datasets from multiple domains, denoted as *source domains*, with similar attributes or features, such as the problem and constraints.

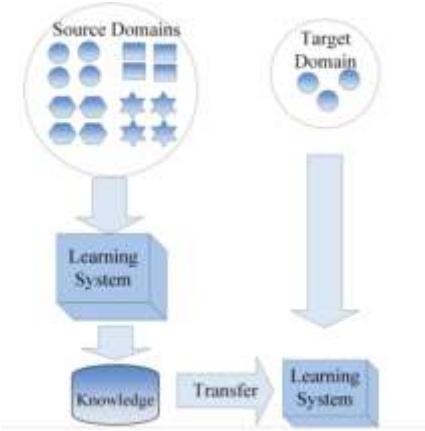

Fig. 7. Transfer Learning.

The distinguishing characteristic of transfer learning from other traditional ML approaches is fact that the training set does not necessarily come from the same domain as the testing set. Moreover, it can train on data from several domains individually or combined together using regular or adapted machine learning algorithms

5) Lifelong Learning

Lifelong learning mimics human learning; learning is continuous; knowledge is retained and used to solve different problems. It is directed to maximize overall learning, to be able to solve a new task by training either on one single domain or on heterogeneous domains collectively . The learning outcomes from the training process are collected and combined together in a space called the *topic model* or *knowledge model*. In the case of training on heterogeneous domains, transfer learning might be used in the combining step to create such a topic model. The existing knowledge in this topic model is used to perform a new task regardless of where the knowledge comes from.

Lifelong learning is related to online learning and transfer learning. Like online learning, lifelong learning is a continuous process; however, whereas online learning considers only a single domain, lifelong learning includes a multitude of domains. Like transfer learning, lifelong learning is capable of transferring knowledge among domains. But, unlike transfer learning, lifelong learning is a continuous process over time because the topic space is refined each time a new learning outcome arrives.

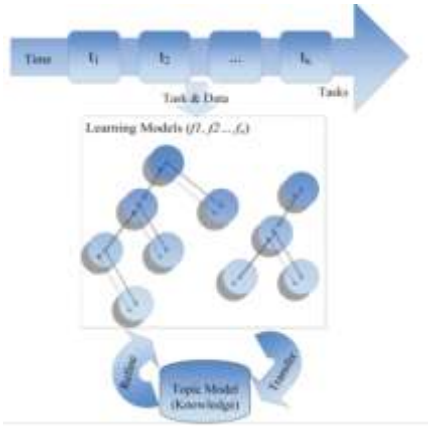

Fig. 8.Transfer Learning.

6) Ensemble Learning

Ensemble learning combines multiple learners to obtain better learning outcomes (e.g., prediction, classification) than those obtained from any constituent learner . Fig. 9 presents an abstract view of the ensemble learning process. Typically, the overall outcome is determined by a voting process among the weighted outcomes individual learners [121]. These individual learners can be similar or from completely different categories, including those belonging to supervised and unsupervised ML. The weighting mechanism assigns a value to each learning output point and combines them..

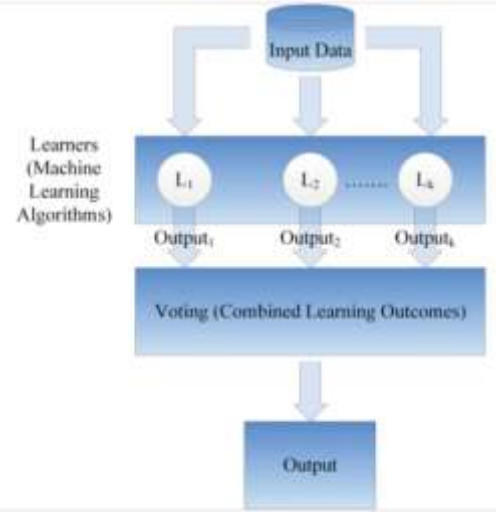

Fig. 9.Ensemble Learning.

There are two main ways to apply ensemble learning: the first one trains different learners, each one on the complete dataset, whereas the second one splits the dataset and trains each learner (same or different) only on a subset. The second approach has potential in the Big Data context because it can speed up and improve the learning process. Basically, improvement is achieved by splitting up large volumes of data into small disjoint datasets

Ensemble learning plays a key role in emphasizing correctness of learning outcomes as well as speeding up the learning process. With respect to correctness, using several different learners or training with different subsets has the potential to minimize the error rates.

Conclusions:

This paper has provided a systematic review of the challenges associated with machine learning in the context of Big Data.The use of the Big Data definition to categorize the challenges of machine learning enables the creation of cause-effect connections for each of the issues. Furthermore, the creation of explicit relations between approaches and challenges enables a more thorough understanding of ML with Big Data. This fulfills the first objective of this work; to create a foundation for a deeper understanding of machine learning with Big Data.

Another objective of this study was to provide researchers with a strong foundation for making easier and better-informed choices with regard to machine learning with Big Data.from the development or adaptation of new machine learning paradigms to tackle unresolved challenges, to the combination

of existing solutions to achieve further performance improvements, this paper has identified research opportunities. This work has therefore accomplished its last objective by providing the academic community with potential directions for future work and will hopefully serve as groundwork for great improvements in the field of machine learning with Big Data.

REFERENCES:

[1] "Deep learning for detection of diabetic eye disease," https://research.googleblog.com/2016/11/deep-learningfor-detection-of-diabetic.html.

[2] I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. The MIT Press, 2016.

[3] S. H. Bach, B. D. He, A. Ratner, and C. Re, "Learning the structure ´ of generative models without labeled data," in ICML, 2017, pp. 273–282.

[4] N. Polyzotis, S. Roy, S. E. Whang, and M. Zinkevich, "Data lifecycle challenges in production machine learning: A survey," SIGMOD Rec., vol. 47, no. 2, pp. 17–28, Jun. 2018.

[5] ——, "Data management challenges in production machine learning," in SIGMOD, 2017, pp. 1723–1726.

[6] "Google cloud automl," https://cloud.google.com/automl/.

[7] "Microsoft custom vision," https://azure.microsoft.com/en-us/ services/cognitive-services/custom-vision-service/.

[8] "Amazon sagemaker," https://aws.amazon.com/sagemaker/.

[9] A. Bhardwaj, A. Deshpande, A. J. Elmore, D. Karger, S. Madden, A. Parameswaran, H. Subramanyam, E. Wu, and R. Zhang, "Collaborative data analytics with datahub," PVLDB, vol. 8, no. 12, pp. 1916–1919, Aug. 2015. 17

[10] A. P. Bhardwaj, S. Bhattacherjee, A. Chavan, A. Deshpande, A. J. Elmore, S. Madden, and A. G. Parameswaran, "Datahub: Collaborative data science & dataset version management at scale," in CIDR, 2015.

[11] S. Bhattacherjee, A. Chavan, S. Huang, A. Deshpande, and A. Parameswaran, "Principles of dataset versioning: Exploring the recreation/storage tradeoff," PVLDB, vol. 8, no. 12, pp. 1346– 1357, Aug. 2015.

[12] A. Y. Halevy, "Data publishing and sharing using fusion tables," in CIDR, 2013.

[13] H. Gonzalez, A. Y. Halevy, C. S. Jensen, A. Langen, J. Madhavan, R. Shapley, and W. Shen, "Google fusion tables: data management, integration and collaboration in the cloud," in SoCC, 2010, pp. 175–180.

[14] H. Gonzalez, A. Y. Halevy, C. S. Jensen, A. Langen, J. Madhavan, R. Shapley, W. Shen, and J. Goldberg-Kidon, "Google fusion tables: web-centered data management and collaboration," in SIGMOD, 2010, pp. 1061–1066.

[15] "Ckan," http://ckan.org.

[16] "Quandl," https://www.quandl.com.

[17] "Datamarket," https://datamarket.com.

[18] "Kaggle," https://www.kaggle.com/.

[19] I. G. Terrizzano, P. M. Schwarz, M. Roth, and J. E. Colino, "Data wrangling: The challenging yourney from the wild to the lake," in CIDR, 2015.

[20] A. Halevy, F. Korn, N. F. Noy, C. Olston, N. Polyzotis, S. Roy, and S. E. Whang, "Goods: Organizing google's datasets," in SIGMOD, 2016, pp. 795–806.

[21] R. Castro Fernandez, D. Deng, E. Mansour, A. A. Qahtan, W. Tao, Z. Abedjan, A. Elmagarmid, I. F. Ilyas, S. Madden, M. Ouzzani, M. Stonebraker, and N. Tang, "A demo of the data civilizer system," in SIGMOD, 2017, pp. 1639–1642.

[22] D. Deng, R. C. Fernandez, Z. Abedjan, S. Wang, M. Stonebraker, A. K. Elmagarmid, I. F. Ilyas, S. Madden, M. Ouzzani, and N. Tang, "The data civilizer system," in CIDR, 2017.

[23] Y. Gao, S. Huang, and A. G. Parameswaran, "Navigating the data lake with DATAMARAN: automatically extracting structure from log datasets," in SIGMOD, 2018, pp. 943–958.

[24] M. J. Cafarella, A. Halevy, D. Z. Wang, E. Wu, and Y. Zhang, "Webtables: Exploring the power of tables on the web," PVLDB, vol. 1, no. 1, pp. 538–549, 2008.

[25] M. J. Cafarella, A. Y. Halevy, H. Lee, J. Madhavan, C. Yu, D. Z. Wang, and E. Wu, "Ten years of webtables," PVLDB, vol. 11, no. 12, pp. 2140–2149, 2018.

[26] "Google dataset search," https://www.blog.google/products/ search/making-it-easier-discover-datasets/.