20BCE1025
Abhishek N N

CSE3505 FOUNDATIONS OF DATA ANALYTICS DA_1

Problem: Big Data Analytics through Machine Learning

Research Papers:

1) Hariri, R. H., Fredericks, E. M., & Bowers, K. M. (2019). Uncertainty in big data analytics: survey, opportunities, and challenges. Journal of Big Data, 6(1), 1-16.

2) Salkuti, S. R. (2020). A survey of big data and machine learning. International Journal of Electrical & Computer Engineering (2088-8708), 10(1).

3) Wang, L., & Alexander, C. A. (2016). Machine learning in big data. International Journal of Mathematical, Engineering and Management Sciences, 1(2), 52-61.

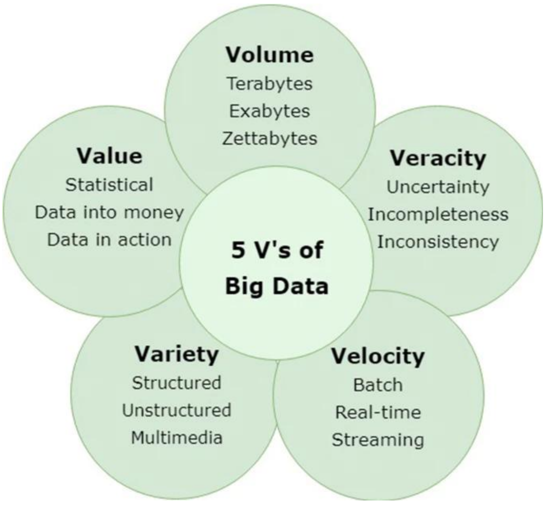
4) L'heureux, A., Grolinger, K., Elyamany, H. F., & Capretz, M. A. (2017). Machine learning with big data: Challenges and approaches. Ieee Access, 5, 7776-7797.

5) Qiu, J., Wu, Q., Ding, G., Xu, Y., & Feng, S. (2016). A survey of machine learning for big data processing. EURASIP Journal on Advances in Signal Processing, 2016(1), 1-16.

Problem:

Big data analytics has gained wide attention from both academia and industry as the demand for understanding trends in massive datasets increases. Recent developments in sensor networks, cyber-physical systems, and the ubiquity of the Internet of Things (IoT) have increased the collection of data (including health care, social media, smart cities, agriculture, finance, education, and more) to an enormous scale. However, the data collected from sensors, social media, financial records, etc. is inherently uncertain due to noise, incompleteness, and inconsistency. The analysis of such massive amounts of data requires advanced analytical techniques for efficiently reviewing and/or predicting future courses of action with high precision and advanced decision-making strategies. As the amount, variety, and speed of data increases, so too does the uncertainty inherent within, leading to a lack of confidence in the resulting analytics process and decisions made thereof.

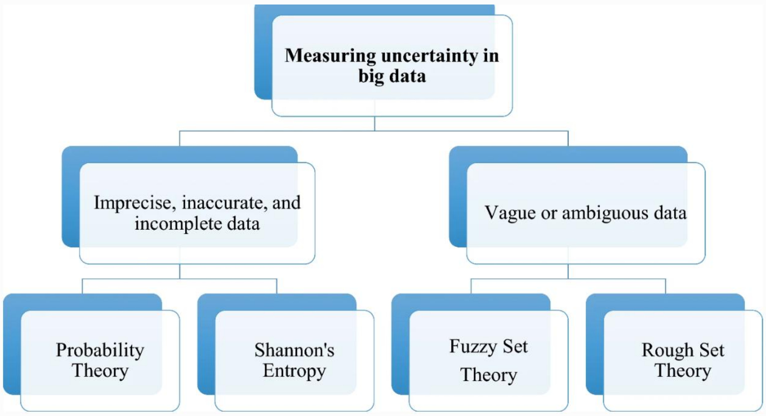
Common big data characteristics



Volume refers to the massive amount of data generated every second and applies to the size and scale of a dataset.
Variety refers to the different forms of data in a dataset including structured data, semi-structured data, and unstructured data.
Velocity comprises the speed (represented in terms of batch, near-real time, real time, and streaming) of data processing, emphasizing that the speed with which the data is processed must meet the speed with which the data is produced.
Veracity represents the quality of the data (e.g., uncertain or imprecise data).
Value represents the context and usefulness of data for decision making, whereas the prior V's focus more on representing challenges in big data.

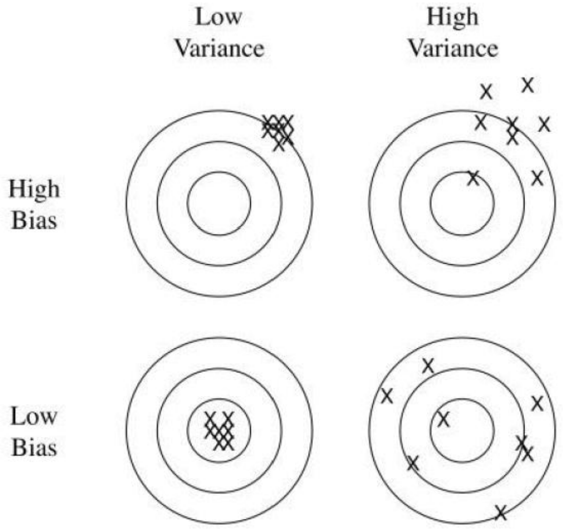
Uncertainty

Generally, "uncertainty is a situation which involves unknown or imperfect information". Uncertainty exists in every phase of big data learning and comes from many different sources, such as data collection (e.g., variance in environmental conditions and issues related to sampling), concept variance (e.g., the aims of analytics do not present similarly) and multimodality (e.g., the complexity and noise introduced with patient health records from multiple sensors include numerical, textual, and image data).



Variance and Bias

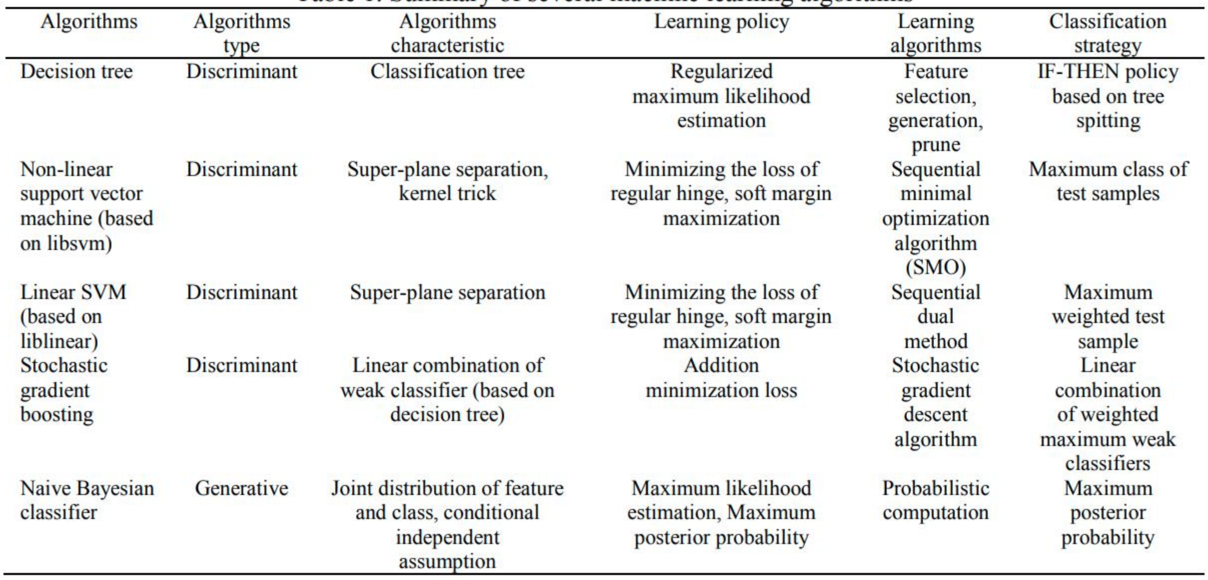
Machine learning relies upon the idea of generalization; through observations and manipulations of data, representations can be generalized to enable analysis and prediction. Generalization error can be broken down into two components: variance and bias [45]: Fig. 2 illustrates the relationship between them. *Variance* describes the consistency of a learner's ability to predict random things, whereas *bias* describes the ability of a learner to learn the wrong thing [37]. Ideally, both the variance and the bias error should be minimized to obtain an accurate output. However, as the volume of data increases, the learner may become too closely biased to the training set and may be unable to generalize adequately for new data. Therefore, when dealing with Big Data, caution should be taken as bias can be introduced, compromising the ability to generalize.



Solution:

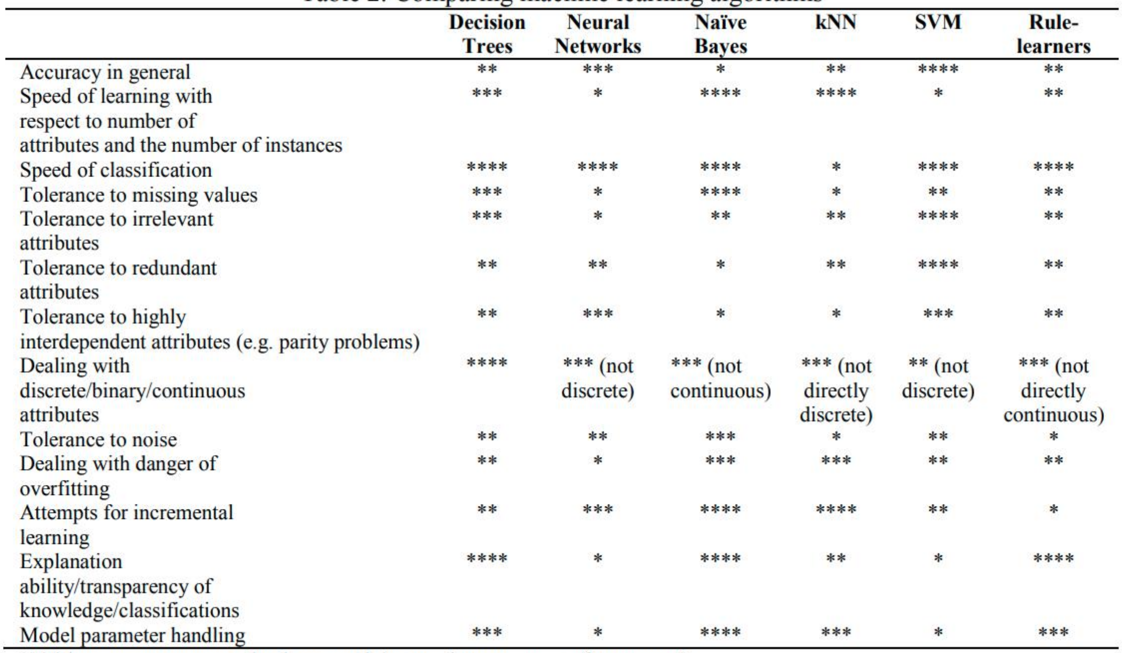
Current algorithms and methods:

Table 1. Summary of several machine learning algorithms



Algorithms	Algorithms type	Algorithms characteristic	Learning policy	Learning algorithms	Classification strategy
Decision tree	Discriminant	Classification tree	Regularized maximum likelihood estimation	Feature selection, generation, prune	IF-THEN policy based on tree spitting
Non-linear support vector machine (based on libsvm)	Discriminant	Super-plane separation, kernel trick	Minimizing the loss of regular hinge, soft margin maximization	Sequential minimal optimization algorithm (SMO)	Maximum class of test samples
Linear SVM (based on liblinear)	Discriminant	Super-plane separation	Minimizing the loss of regular hinge, soft margin maximization	Sequential dual method	Maximum weighted test sample
Stochastic gradient boosting	Discriminant	Linear combination of weak classifier (based on decision tree)	Addition minimization loss	Stochastic gradient descent algorithm	Linear combination of weighted maximum weak classifiers
Naive Bayesian classifier	Generative	Joint distribution of feature and class, conditional independent assumption	Maximum likelihood estimation, Maximum posterior probability	Probabilistic computation	Maximum posterior probability

Table 2. Comparing machine learning algorithms



	Decision Trees	Neural Networks	Naïve Bayes	kNN	SVM	Rule-learners
Accuracy in general	**	***	*	**	****	**
Speed of learning with respect to number of attributes and the number of instances	***	*	****	****	*	**
Speed of classification	****	****	****	*	****	****
Tolerance to missing values	***	*	****	*	**	**
Tolerance to irrelevant attributes	***	*	**	**	****	**
Tolerance to redundant attributes	**	**	*	**	****	**
Tolerance to highly interdependent attributes (e.g. parity problems)	**	***	*	*	***	**
Dealing with discrete/binary/continuous attributes	****	*** (not discrete)	*** (not continuous)	*** (not directly discrete)	** (not discrete)	*** (not directly continuous)
Tolerance to noise	**	**	***	*	**	*
Dealing with danger of overfitting	**	*	***	***	**	**
Attempts for incremental learning	**	***	****	****	**	*
Explanation ability/transparency of knowledge/classifications	****	*	****	**	*	****
Model parameter handling	***	*	****	***	*	***

(**** stars represent the best and * star the worst performance)

APPROACHES				CHALLENGES																	
				VOLUME								VARIETY			VELOCITY				VERACITY		
				Processing Performance	Curse of Modularity	Class Imbalance	Curse of Dimensionality	Feature Engineering	Non-linearity	Bonferonni"s Principle	Variance and Bias	Data locality	Data Heterogeneity	Dirty and noisy Data	Data availability	Real-time Processing/Streaming	Concept drift	I.i.d	Data Provevance	Data Uncertainty	Dirty and Noisy Data
MANIPULATIONS	Data Manipulations	Dimensionality Reduction		✓			✓														
		Instance Selection		✓	✓																
		Data Cleaning												✓							✓
	Processing Manipulations	Vertical Scaling		✓															*		
		Horizontal Scaling	Batch-oriented	✓	✓		*					✓							*		
			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 3. Comparison of Big Data Technologies

	In-memory database	MPP database	Big Data appliance	Hadoop	NoSQL database
Consistent	W	W	W	P	P
Available	W	W	W	P	P
Fault tolerant	W	W	P	W	W
Suitable for real-time transactions	W	W	W	F	F
Suitable for analytics	P	P	W	W	F
Suitable for extremely big data	F	P	P	W	W
Suitable for unstructured data	F	F	P	W	W

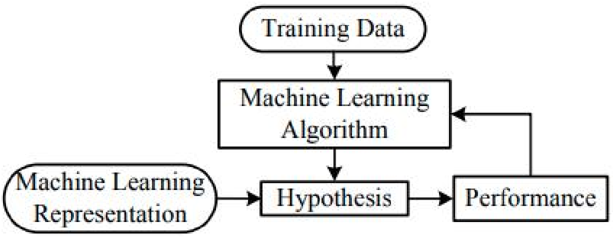
W: Meets widely held expectations.

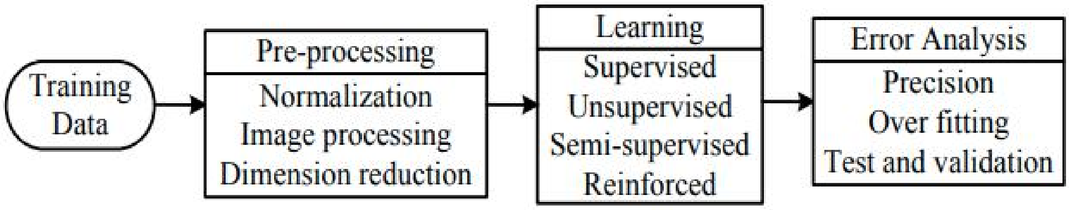
P: Potentially meets widely held expectations.

F: Fails to meet widely held expectations.

General Machine Learning Process:

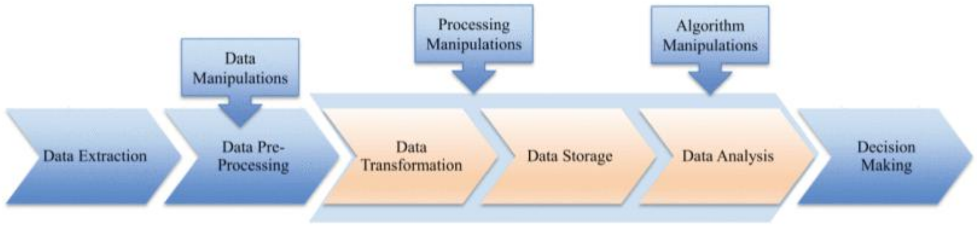
ML is the science which give the computers the ability to learn and predict from the experience without explicitly programmed. If a computer program can improve its performance by learning from previous experience then one can say that it has learned. Machine learning is more closed to data analysis rather than AI. Machine learning uses algorithms that allow computers to iteratively learn from data. In past decades, ML has reached to a new level. ML has given us self-driving car, effective web search, human voice Int J Elec & Comp Eng ISSN: 2088-8708 A survey of big data and machine learning (Surender Reddy Salkuti) 577 recognition, image recognition and many more [20]. Every day we use it several times without knowing it. The process of ML is depicted in Figure as below.

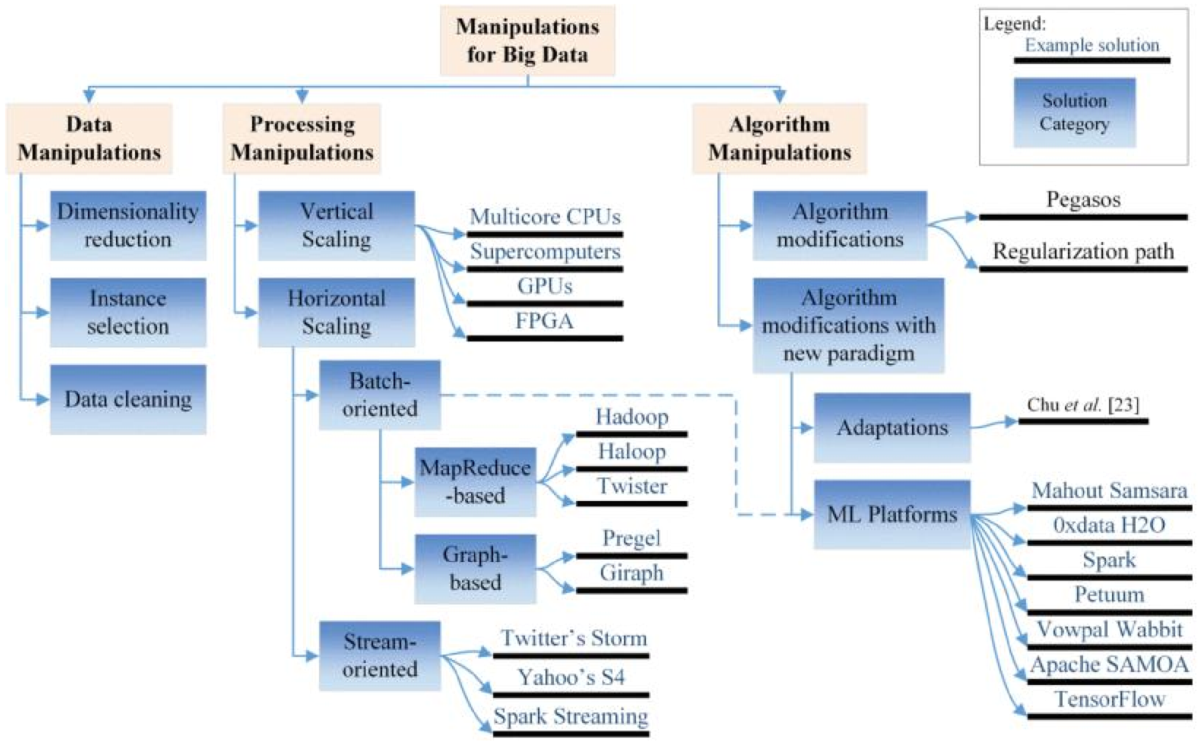




Manipulations for Big Data

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. Below 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 as below. The examples included are only representatives and in no way provide a comprehensive list of solutions.





Machine Learning Paradigms for Big Data

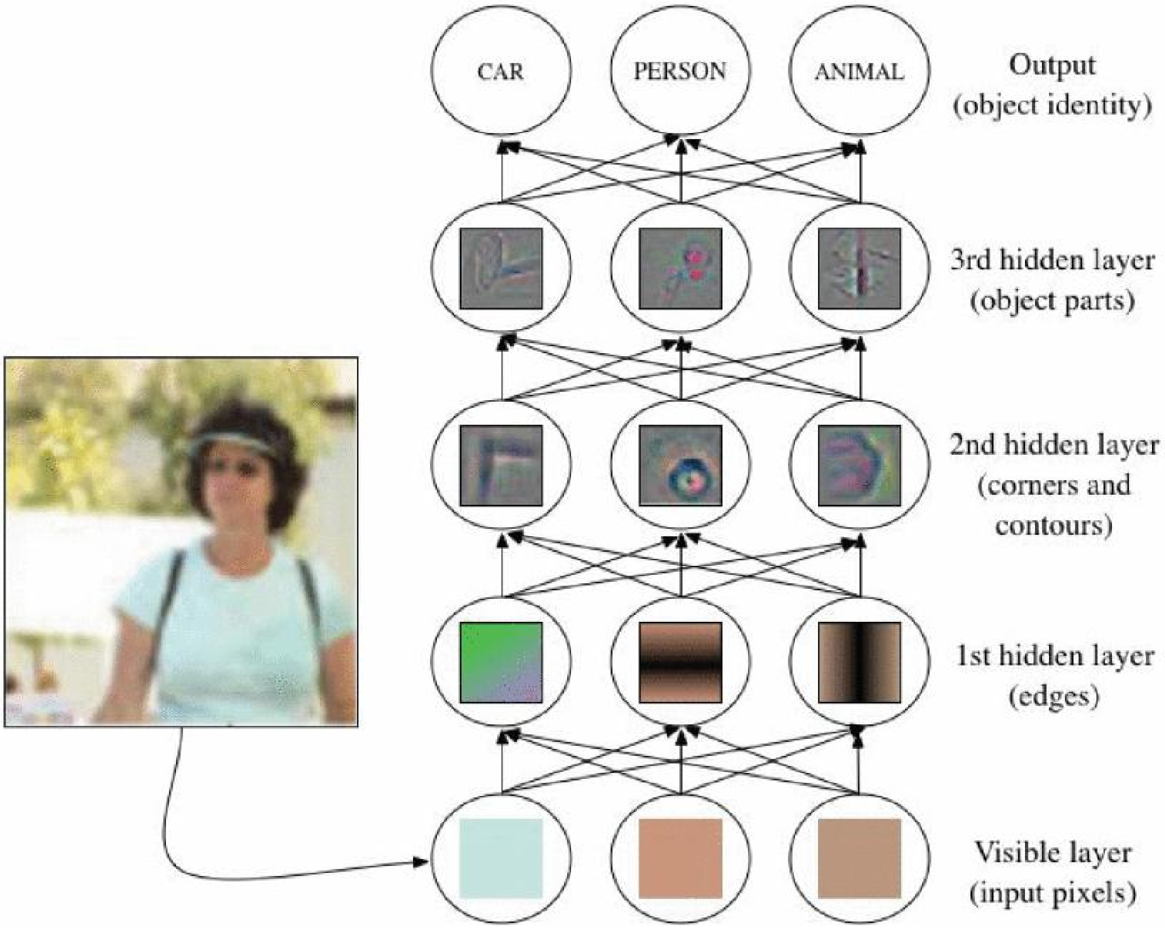
A variety of learning paradigms exists in the field of machine learning; however, not all types are relevant to all areas of research. For example, Deng and Li presented a number of paradigms that were applicable to speech recognition. Congruently, the work presented here includes machine learning paradigms relevant in the Big Data context, along with how they address the identified challenges.

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 [14]. 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. Below provides an abstract view of the deep learning process. Each layer learns a specific feature: edges, corners and contours, and object parts.



Observation and Conclusion:

Here we discussed how uncertainty can impact big data, both in terms of analytics and the dataset itself. also discussed about the state of the art with respect to big data analytics techniques, how uncertainty can negatively impact such techniques, and examine the open issues that remain. For each common technique, we have summarized relevant research to aid others in this community when developing their own techniques. We have discussed the issues surrounding the five V's of big data, however many other V's exist. In terms of existing research, much focus has been provided on volume, variety, velocity, and veracity of data, with less available work in value (e.g., data related to corporate interests and decision making in specific domains).

A detailed analysis of big data and machine learning (ML) has been presented here. Big data analytics involves the processes of searching a database, mining, and analysing data dedicated to improve the performance of the company. ML focuses on the development of computer programs that can teach themselves to grow and change when exposed to the new data. Applications of big data and ML in various industries such as electrical power and energy including smart grid, transportation, health care, education, e-commerce, financial services, marketing and sales, etc. Various challenges and opportunities related to big data and machine learning can also be infered from here.

Splitting criteria of decision trees are chosen based on some quality measures, which requires handling the entire data set of each expanding nodes. This makes it difficult for decision trees to be used in big data applications.

SVM shows very good performance to data sets in a moderate size. It has inherent limitations to big data applications.

Deep learning is suited to address issues related to volume and variety of big data. However, it has some restrictions in big data because it requires much training time. PLANET can deal with big volume of data, but is not applicable to data with categorical attributes. Machine learning applications in big data has met challenges such as memory constraint, no support (in iterations) from MapReduce, difficulty in dealing with big data due to Vs (such as high velocity, volume, and variety, etc.), and learning training limited to a certain number of class types or a particular labeled datasets, etc. Some technology progress has been made such as faceted learning for hierarchical data structure, multi-task learning in in parallel, multi-domain/ cross-domain representation-learning, streaming data processing, high-dimensional data processing, and online feature selection, etc. These areas and the above challenges about machine learning in big data also can be further research topics.

Here we provided a systematic review of the challenges associated with machine learning in the context of Big Data and categorized them according to the V dimensions of Big Data. Moreover, it has presented an overview of ML approaches and discussed how these techniques overcome the various challenges identified.

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.

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, here we identified research opportunities.

Big data are now rapidly expanding in all science and engineering domains. Learning from these massive data is expected to bring significant opportunities and transformative potential for various sectors. However, most traditional machine learning techniques are not inherently efficient or scalable enough to handle the data with the characteristics of large volume, different types, high speed, uncertainty and incompleteness, and low value density. In response, machine learning needs to reinvent itself for big data processing. Here we began with a brief review of conventional machine learning algorithms, followed by several current advanced learning methods. Then, a discussion about the challenges of learning with big data and the corresponding possible solutions in recent researches was given.

The topic is more of open ended with no perticular solution works for all thing, with more scope for immense development and research.