Asian Journal of Science and Applied Technology
ISSN: 2249-0698 (P) Vol.6, No.1, 2017, pp.18-21

DOI: https://doi.org/10.51983/ajsat-2017.6.1.942

Effective Machine Learning Techniques used in Big Data Analytics

S. Senthil Kumar[1] and V.Kathiresan[2]
[1]Assistant Professor, Department of Commerce with Computer Applications,
[2]Head, Department of Computer Applications (PG),
Dr.SNS Rajalakshmi College of Arts and Science (Autonomous), Coimbatore, Tamil Nadu, India
E-mail: ssksnsmca@gmail.com, vkathirmca@gmail.com
(Received 10 February 2017; Revised 28 February 2017; Accepted 21 March 2017; Available online 27 March 2017)

***Abstract* - Big data is a general term for massive amount of digital data being collected from various sources that are too large and raw in form. Big data deals with new challenges like complexity, security, risks to privacy. Big data is redefining the data management from extraction, transformation and processing to cleaning and reducing [1].**

There has been a lot of growth in the amount of data generated by web these days. The data has been so large that it becomes difficult to analyse it with the help of our traditional mining methods. Big data term has been coined for data that exceeds the processing capability [2]. Moreover, the rising data volume has contributed to the increasing demand for big data analytics.

***Keywords:* Big data, Feature Selection, Supervised Learning, Unsupervised Learning, Deep Learning**

I. INTRODUCTION

Big Data is a new term used to identify the datasets that due to their large size and complexity. Big Data are now rapidly expanding in all science and engineering domains, including physical, biological and biomedical sciences. Big Data mining is the capability of extracting useful information from these large datasets. Big data is high-volume, high velocity and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making.

II. THREE MAIN KEY CHARACTERISTICS

1. ***Volume:*** The size of data is now larger than terabytes and petabytes. This large scale makes it difficult to analyse using conventional methods.
2. ***Velocity:*** Big data should be used to mine large amount of data within a pre-defined period of time. The traditional methods of mining may take huge time to mine such a volume of data.
3. ***Variety*:** Big data comes from various sources. It is designed to handle structured, semi-structured as well as unstructured data. Whereas the traditional methods were designed to handle structured data and that too not of such large volume.

III.APPLICATIONS OF BIG DATA

a. In social networking sites to find for usage patterns
b. In Google search
c. Astronomy
d. Sensor networks
e. Government data
f. Web logs
g. Mobile phones
h. Natural disaster and resource management
i. Scientific research

IV. FEATURES OF BIG DATA ARE

1. It is huge in size.
2. The data keep on changing time time to time.
3. Its data sources are from different phases.
4. It is free from the influence, guidance, or control of anyone.
5. It is too much complex in nature, thus hard to handle.

Due to largeness in size, decentralized control and different data sources with different types the Big Data becomes much complex and harder to handle. We cannot manage them with the local tools those we use for managing the regular data in real time. For major Big Data-related applications, such as Google, Flicker, Facebook, a large number of server farms are deployed all over the world to ensure nonstop services and quick responses for local markets. The best example of big data Facebook, lots of numbers of people are uploading their data in various types such as text, images or videos.

V. BIG DATA ANALYTICS

Big Data Analytics is aimed at making sense of data by applying efficient and scalable algorithms on Big Data for its analysis, learning, modelling, visualization and understanding. This includes the design of efficient and effective algorithms and systems to integrate the data and uncover the hidden values from data. It also includes methodologies and algorithms for automatic or mixed-initiative knowledge discovery and learning, data transformation and modelling, predictions and explanations of the data. Breakthroughs in this area include new algorithms, methodologies, systems and applications for knowledge discovery, understanding and applications based on the Big Data. New computing paradigms are expected in new areas such as human computation, crowd sourcing,

sentiment analysis as well as data visualization technologies.

VI. TECHNIQUES FOR BIG DATA ANALYTICS

Supervised, unsupervised, and hybrid machine learning approaches are the most widely used tools for descriptive and predictive analytics on big data. The problem of big data volume can be somewhat minimized by dimensionality reduction.

Another important tool used in big data analytics is mathematical optimization. Subfields of optimization, such as constraint satisfaction programming, dynamic programming, and heuristics & meta heuristics are widely used in AI and machine learning problems. Other important optimization methods include multi-objective and multi-modal optimization methods, such as pareto optimization and evolutionary algorithms , respectively.

Big data analytics has a close proximity to data mining approaches. Mining big data is more challenging than traditional data mining due to massive data volume. The common practice is to extend the existing data mining algorithms to cope with massive datasets, by executing on samples of big data and then merging the sample results. This kind of clustering algorithms include CLARA (Clustering LARge Applications) [3] and BIRCH (Balanced Iterative Reducing using Cluster Hierarchies) [4].

VII. MACHINE LEARNING TECHNIQUES FOR BIG DATA

They have been found very effective and relevant to many real world applications in bioinformatics, network security, healthcare, banking and finance and transportations. Over time, bioinformatics and health related data are created and accumulated continuously, resulting in an incredible volume of data. Newer forms of big data, such as 3D imaging, genomics and biometric sensor readings are also fuelling this exponential growth. Future applications of real-time data, such as early detection of infections/diseases and fast application of the appropriate treatments (not just broad-spectrum antibiotics) could reduce patient morbidity and mortality. Already, real time streaming data monitors neonates in the ICU, catching life-threatening infections at real time. The ability to perform real-time analytics against such voluminous stream data across all specialties would revolutionize healthcare. There lies data with volume, velocity, and variety.

Machine learning is a field of computer science that studies the computational methods that learn from data [6].There are mainly two types of learning methods in machine learning, viz., supervised and unsupervised learning methods[7]. In supervised learning, a method learns from a set of objects with class label, often called a training set.

Fig.1 Human Vs Machine Learning

The acquired knowledge is used to assign label to unknown objects often called test objects. On the other hand, unsupervised learning methods do not depend on the availability of prior knowledge or training instances with class labels. All these machine learning methods require pre processing of datasets for effective results. Feature selection is one of the important pre processing tasks that leads to improved result and reduced time requirement. Hybrid learning methods, such as Deep learning, have become popular in the recent years and provide significantly high accuracy.

A.Feature selection

The main objective of feature selection is to select a subset of most relevant and non-redundant features that can increase the performance of a learning method. A feature selection method can improve the performance of prediction models by removing irrelevant and redundant features with alleviating the effect of the curse of dimensionality, enhancing the generalization performance, speeding up the learning process, and improving the model interpretability.

A feature selection plays a major role in identifying the most important features from a ultrahigh dimensional big dataset. The selected feature set can be used for processing large volume of data to take instant decision in short period of time. Especially, in big data analytics, relevant features can be selected from large data using both supervised learning as well as unsupervised learning. Hence, ranking the features based on their relevance and selecting the most relevant features can vastly improve the generalization performance.

B.Supervised learning

In supervised learning, labeled training examples are used to train the learning algorithm. The objective of a supervised learning model is to predict the class labels of test instances based on knowledge gained from the available training instances. Within supervised learning family we can further distinguish between classification models which focus on prediction of discrete (categorical) outputs or regression models which predict continuous outputs. Among large number of models reported in the literature linear and nonlinear density-based classifiers, decision trees, naive Bayes, support vector machines (SVMs), neural networks and K-nearest neighbour (KNN) are the most frequently used methods in many applications.

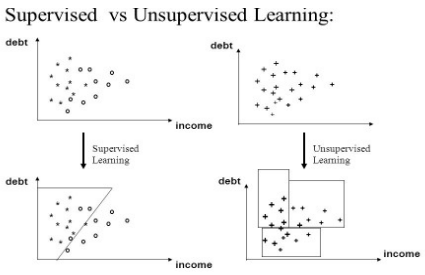


Fig.(b) Supervised learning Vs Unsupervised learning

In Big Data analytics, we need some advanced supervised approaches for parallel and distributed learning such as Multi-hyperplane Machine (MM) classification model [8], divide-and-conquer SVM [9], and neural network classifiers. Among these SVM is one of the most efficient and widely used supervised learning method and several modified SVM methods have been introduced for big data analytics.

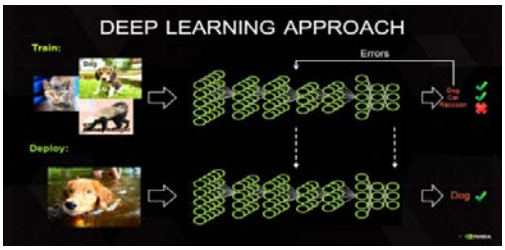
C.Unsupervised learning

Unsupervised learning do not use the class labels of the objects for learning . Clustering is an unsupervised technique that attempts to group objects to optimize the criterion that states that distance among objects in the same cluster is minimized and distance among objects in different clusters is maximized . A major issue in clustering is the computation of distance between a pair of objects. Various proximity measures have been used for this purpose, such as Euclidean, Cosine, and city block distance. In traditional clustering, all the features are used while computing the distance between a pair of objects. A cluster is a group of objects that are close to each other with respect to their mutual distance. In other words, they are similar in nature over the entire set of features. Triclustering operates on such datasets to generate triclusters. A tricluster is a group of objects that are not only similar over a subset of features, but are also similar across a subset of time points [10]. Triclustering promotes grouping of objects, features and time points simultaneously.

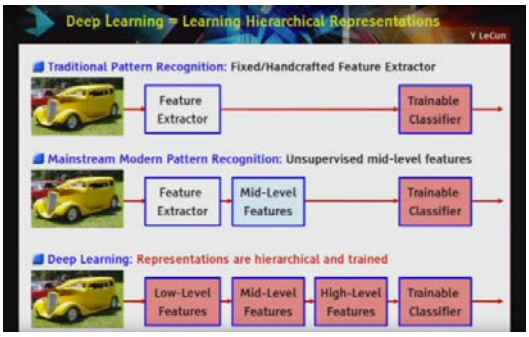
Soft computing-based clustering methods use soft computation tools, such as fuzzy set and neural network. Fuzzy c-means [11] a soft computing-based clustering method, is a crisp method, which allows objects to belong to more than one cluster with the constraint that the sum of membership of an object across all the clusters is equal to one. This method tries to find a crisp partition that minimizes a cost function.

D.Deep Learning

Deep learning attempts to model high-level abstractions in data using supervised and/or unsupervised learning algorithms, in order to learn from multiple levels of abstractions. It uses hierarchical representations of data for classification. Deep learning methods have been used in many applications, viz., pattern recognition, computer vision, natural language processing and speech recognition exponential increase of data in these applications, deep learning is useful for accurate prediction from voluminous data.



In recent years, researchers have developed effective and scalable parallel algorithms for training deep models Many organizations use deep learning for decision making, information retrieval, and semantic indexing.



VIII. CONCLUSION

A deep learning architecture is shown in Input data are partitioned into multiple samples for data abstractions. The intermediate layers are used to process the features at multiple levels for prediction from data. The final prediction is performed at the output layer using the outputs of its immediate upper layer. Deep learning represents data in multiple layers. It can efficiently process high volume of data, where shallow learning fails to explore due to the complexities of data patterns. Moreover, deep learning is quiet suitable for analyzing unstructured and heterogeneous data collected from various sources. Traditional neural networks pose two problems, viz., poor performance due to local optima of a non-convex object function and incapability to exploit unlabeled data, which are abundant and cheap. To overcome these limitations of traditional neural networks, Deep Belief Networks (DBN) [12] was introduced with a deep learning architecture to learn from both labeled and unlabeled data.

REFERENCES

[1] Gang-Hoon Kim, Silvana Trimi, Ji-Hyong Chung, "Big-Data Applications in the Government Sector", Communications of the ACM, Vol. 57, No.3, Pages 78-75

[2] Richa Gupta, Sunny Gupta, Anuradha Singhal, "Big Data: Overview", International Journal of Computer Trends and Technology (IJCTT), Vol 9, No.5, March 2014

[3] L. Kaufman and P. J. Rousseeuw, “Finding groups in data. an introduction to cluster analysis,” Wiley Series in Probability and Mathematical Statistics. Applied Probability and Statistics, New York: Wiley, 1990, vol. 1, 1990.

[4] T. Zhang, R. Ramakrishnan, and M. Livny, "Birch: an efficient data clustering method for very large databases," in ACM SIGMOD Record, Vol. 25, No. 2. ACM, 1996, pp. 103–114.

[5] T. B¨ack, "Evolutionary computation: Toward a new philosophy of machine intelligence," 1997.

[6] C. M. Bishop et al., Pattern recognition and machine learning. springer New York, 2006, Vol. 4, No. 4.

[7] D. K. Bhattacharyya and J. K. Kalita, Network anomaly detection: A machine learning perspective. CRC Press, 2013.

[8] N. Djuric, "Big data algorithms for visualization and supervised learning," Ph.D. dissertation, Temple University, 2014.

[9] C.-J. Hsieh, S. Si, and I. S. Dhillon, "A divide-and-conquer solver for kernel support vector machines," arXiv preprint arXiv:1311.0914, 2013

[10] H. Ahmed, P. Mahanta, D. Bhattacharyya, J. Kalita, and A. Ghosh, "Intersected coexpressed subcube miner: An effective triclustering algorithm," in Information and Communication Technologies (WICT), 2011 World Congress on. IEEE, 2011, pp. 846–851.

[11] F. Hoppner, Fuzzy cluster analysis: methods for classification, data analysis and image recognition. John Wiley & Sons, 1999.

[12] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," Science, Vol. 313, No. 5786, pp. 504–507,2006.