



## **A Study on Machine Learning in Big Data**

**L.DHANAPRIYA and DR. S. MANJU**

Department of Computer Science Sri Ramakrishna College of Arts and  
Science for Women, Coimbatore – 641044, India.

### **Abstract**

In the recent development of IT technology, the capacity of data has surpassed the zettabyte, and improving the efficiency of business is done by increasing the ability of predictive through an efficient analysis on these data which has emerged as an issue in the current society. Now the market needs for methods that are capable of extracting valuable information from large data sets. Recently big data is becoming the focus of attention, and using any of the machine learning techniques to extract the valuable information from the huge data of complex structures has become a concern yet an urgent problem to resolve. The aim of this work is to provide a better understanding of this Machine Learning technique for discovering interesting patterns and introduces some machine learning algorithms to explore the developing trend.



### **Article History**

Received: 09 September  
2017  
Accepted: 19 September  
2017



### **Keywords**

Big Data,  
Machine Learning,  
Machine Learning  
technique, Machine  
Learning algorithms,  
Traditional algorithms.


### **Introduction**

In recent years, as cloud computing, mobile Internet, and Internet of Things rapidly increases the data exponentially. How to learn and deal with these structured, semi-structured and unstructured large-scale massive data as well as the abilities of quickly acquiring valuable information becomes a problem worthy of attention yet of urgent solution<sup>1</sup>. Machine learning is a core research area of artificial intelligence, whose theme is to imitate human learning activities. Machine Learning is concerned with the design and development of algorithms<sup>11</sup>. It includes Supervised Learning, Unsupervised Learning and Semi-Supervised Learning.

Due to the big volume and complexity of big data, Traditional analytics tools are not well suited for capturing the full value of big data. The volume of the captured data is large for comprehensive or any kind of analysis and the range of potential correlations and relationships between disparate data sources from back end databases to live web based clickstreams are too great for any analyst to test all hypotheses and derive all the value buried in the customer data base. The analytical methods which are used in business intelligence and enterprise reporting tools reduce to reporting sums, counts, simple averages and running SQL queries<sup>2</sup>. Online analytical processing is purely an extension of these

**CONTACT** L. Dhanapriya  [ghanapriyabca@gmail.com](mailto:ghanapriyabca@gmail.com)  Department of Computer Science Sri Ramakrishna College of Arts and Science for Women, Coimbatore – 641044, India.

© 2017 The Author(s). Published by Enviro Research Publishers

This is an  Open Access article licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License (<https://creativecommons.org/licenses/by-nc-sa/4.0/>), which permits unrestricted NonCommercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

To link to this article: <http://dx.doi.org/10.13005/ojcs/10.03.15>

basic analytics that still rely on a human to direct activities specifies what should be calculated and non-calculated. Machine learning is ideal to utilize the connection hidden in big data.

Machine learning is a well-known research area and it is mainly concerned with the discovery of models and patterns in data.

**Machine Learning Approaches**

Machine Learning research is focusing on Learning and recognizing complex patterns and to make intellectual decisions based on data<sup>11</sup>. Approaches in Machine learning can be broadly classified into two types:

**Symbolic approaches**

Inductive learning of symbolic descriptions, such as decision trees or logical representations

**Statistical approaches**

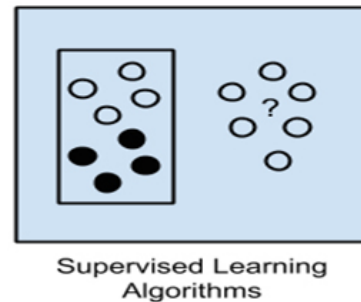
Statistical or pattern-recognition methods, including k-nearest neighbor or instance-based learning Bayesian classifiers neural network learning and support vector.

The machine learning in big data pays closer attention to the study of methods and algorithms.

learning algorithms is useful because it helps in thinking about the roles of the input data and the model preparation process and selecting one of the most appropriate for the problem in order to get the best result<sup>4</sup>. Different learning styles in machine learning algorithms are:

**Supervised Learning**

Here the Input data is called as training data and has a known label or result such as spam/not-spam or a stock price at a time. This model is prepared through a training process in which it is required to make predictions and is corrected when those predictions are wrong<sup>9</sup>. The process continues until the model achieves an expected level of accuracy on the training data. Example problems are classification and regression.



**Unsupervised Learning**

Here the Input data is not labeled and does not have a known result. This model is prepared by deducing structures present in the input data. This may be to extract general rules. To reduce redundancy or to organize data by similarity mathematical process can be used. Few related problems are clustering, dimensionality reduction and association rule learning.

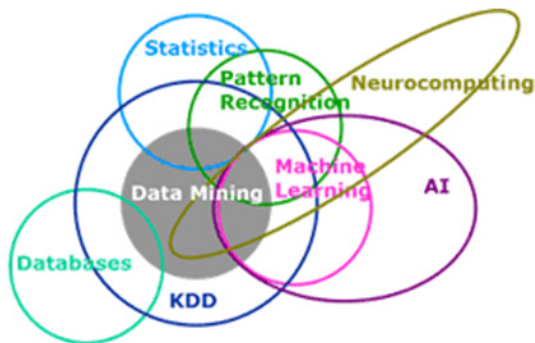
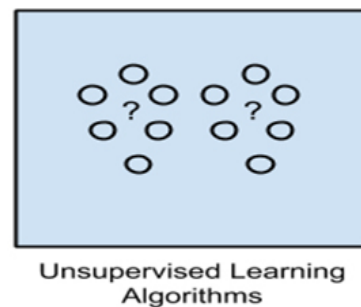


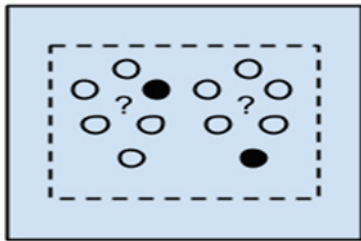
Fig. 1: Machine Learning and Other Research Methods

**Learning Styles In Machine Learning Algorithms**

There are several ways an algorithm can model a problem based on its interaction with the experience or environment or in whatever way the input data can be. The taxonomy or way of organizing machine

**Semi-Supervised Learning**

Here the Input data is a mixture of labeled and unlabeled examples. There is a desired prediction problem but the model must learn the structures to organize the data as well as make predictions. Semi-supervised learning falls between unsupervised learning and supervised learning<sup>12</sup>. Example problems are classification and regression.



Semi-supervised Learning Algorithms

**Machine Learning Algorithms Under Big Data**

Big data has low value density, thus complete set of sample data is frequently adopted when analyzing, which means huge volume of data bring about unprecedented challenges to machine learning. Let's see some machine learning algorithms:

**Decision Tree**

A traditional decision tree algorithm, such as ID3, C4.5 and CART etc. usually follows the greedy top-down method. The core problem of decision tree is choosing the splitting property and its pruning. Traditional decision tree, as a classic classification algorithm, has a problem of large memory consumption when dealing with big data. The advanced method of constructing decision tree from large-scale data to address some limitations in current algorithms, using all the data in training sets but not saving them in the memory. Experiments show that this method calculates faster in large-scale dealing. However, decision trees can be faster when compared to linear classifiers<sup>10</sup>.

**Artificial Neural Networks (ANN)**

Artificial neural network provides a popular and practical method and learn from the sample values for real, discrete or vector function<sup>9</sup>. ANN learning has a good fitting effect for the training data. Different models have been proposed during the research of ANN, they differ mainly in research

approaches, network structure, operating mode, learning algorithms and their respective application. Neural network algorithm is based on empirical risk minimization, having some inherent defects, such as difficulty determinable layer and neuron number, being inclined to fall into local minimum and over fitting phenomenon, which could be well solved in SVM algorithms<sup>6</sup>.

**Support Vector Machine (SVM)**

Support Vector Machine (SVM) has a relatively better performance index<sup>4</sup>, which is based on statistical learning theories. Via learning algorithm, the SVM can automatically pick out support vectors with better distinguishing capability of classifying, constructing out classifiers that maximize intervals between classes, and thus owning better adaptability and efficiency of distinguishing. The goal of SVM algorithm is finding a hyper plane  $H(d)$ , which separates the data in the largest distance to the class field boundary perpendicularly. Thus SVM algorithm is also known as Maximum Margin algorithm. SVM algorithm ultimately eventuates to solving quadratic programming problem, and frequently used SVM algorithms comprise SVM-light, SMO, Chunking, etc.

Traditional statistical machine learning methods are applied to big data classification which involves two bottlenecks: a. computation-intensiveness, hardly applied to large-scale data sets; b. the unknown prediction. As the accuracy of SVM classification algorithm concerns the feature number and the size of the data set, conducting feature selection before classification is beneficial to improve the accuracy.

**Association Rule Algorithm**

Association rules compose no demand on data distribution, and the result is based on data without any subjective assumption objectively reflecting the nature of data. Therefore, association rules have been widely applied to various fields. Association rule algorithm is a solving process from input to output end.

Parallelization and increment are the two ways of solving the association analysis. For parallelization, parallel Apriority algorithm based on Map Reduce is used, whose main operation is to produce candidate item sets, parallelizing the process of producing

the candidate item sets, improving the operating efficiency with better speedup ratio and scalability<sup>7</sup>. Increment is mainly manifested in sequential pattern mining<sup>3</sup>. Advances an Incremental Sequence Mining (ISM) algorithm based on SPADE, which can not only maintain the frequent sequence during the database updating but also provide a user interactive interface to modify restrictions. Machine learning algorithms also include Bayes Algorithm, EM Algorithm, Boosting and Baging Algorithm etc,

### Future Trends

The Developing trend of Machine Learning under Big Data has Ensemble Learning, Transfer Learning, Parallelization and Distribution<sup>5</sup>. The intensive development and increased usage of data mining in specific domain areas, such as bioinformatics, multimedia, text and web data analysis is growing. More tools will emerge in upcoming years for analyzing the Big Data and half of all business analytics software will include the intelligence where it's needed by 2020. On the other hand, data mining can be used for building surveillance systems, recent

research also concentrates on developing algorithms for mining databases without compromising sensitive information.

### Conclusion

So far the research community addresses new open problems and areas, for which data mining is able to provide value-added answers and meaningful results. Big data is the term for a collection of complex data sets, Data mining is an analytic process designed to explore data in search of consistent patterns and then applying the findings to detect the related or new patterns in those new subsets of data. In Recent years, big data is an emerging trend in the world and the need for Big Data mining is raising in all most all domains. With Big data technologies, we will hopefully be able to get enhanced insight and provide wiser and accurate social sensing feedback to better understand our society at real time and machine learning will be a necessary element for data preparation and predictive analysis in businesses moving forward.

### References

- 1 J.W. Han, K. Micheline. *Data Mining: Concepts and Techniques*, Vol. 2, Morgan Kaufmann Publisher (2006)
- 2 N. Marz and J. Warren. *Big Data: Principles and best practices of scalable realtime data systems*. Manning Publications, 2013.
- 3 Schroeder WJ, Zarge JA. Lorensen WE. Decimation of triangle meshes. *Computer Graphics*, 1992.**26**(2):65-70.
- 4 K. Wagstaff. *Machine learning that matters*. In ICML. icml.cc / Omnipress, 2012.
- 5 Manyika J. Chui M. Brown B. et al. Big data: The next frontier for innovation, competition, and productivity [EB/OL]. [http://www.mckinsey.com/insights/business\\_technology/big\\_data\\_the\\_next\\_frontier\\_for\\_innovation](http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation)
- 6 X L Dong, Berti-Equille L, Srivastava D Integrating conflicting data: The role of source dependence. *Proceedings of the VLDB Endowment*. Vol.2 (2009) No.1, p.550-561.
- 7 J.L. Liang, M.H. Zhang and X.Y. Zeng. Distributed Dictionary Learning for Sparse Representation in Sensor Networks. *Image Processing*, IEEE Transactions on Vol.**23** (2014) No.6, p.2528-2541
- 8 Russell, S. J. (2003). *Artificial Intelligence: A Modern Approach* (2nd Edition ed.). Upper Saddle River, NJ, NJ, USA: Prentice Hall.
- 9 Sleeman, D. H. (1983). *Inferring Student Models for Intelligent CAI*. Machine Learning. TiogaPress.
- 10 Mitchell, T. M. (2006). *The Discipline of Machine Learning*. Machine Learning Department technical report CMU-ML-06-108, Carnegie Mellon University.
- 11 Mrs.S.Manju & Dr.M.Punithavalli 2011, "An Analysis of Q-Learning Algorithms to improve the efficiency of Reward function" *International Journal of Computer science & Engineering*, Vol. 3 No. 2 pp: 814-820, ISSN : 2229-5631
- 12 S.Parvathavardhini and Dr. S.Manju "Analysis on Machine Learning Techniques" *International Journal of Computer Sciences and Engineering (IJCSE)*, Vol-4(8), pp 59-77 Aug 2016, E-ISSN: 2347-2693.