



## Research on Machine Learning in the Age of Large Data

**Zhiyang Song**

*School of Computer, Guangdong University of Science & Technology, Dongguan 523083, Guangdong, China*

---

**Abstract**—Large data has the characteristics of sparse attributes, ultra-high dimension, high noise, data drift and complex relationships, which makes it difficult for traditional machine learning algorithms to process and analyze effectively. The arrival of the new era of big data means that there will be a new way to deal with big data, that is, the close relationship between machine learning and big data analysis. Machine learning, which is emerging in the field of artificial intelligence, will combine big data to achieve more powerful functions in more fields, and artificial intelligence technology will also have new breakthroughs. The theory and method of machine learning are studied to realize knowledge discovery in super-high dimensional and sparse large data.

**Keywords**—Machine Learning, Large Data, Explicit Procedures.

---

### I. INTRODUCTION

With the arrival of the era of big data, big data has gradually become a hot spot in academia and industry. International data companies have defined four characteristics of big data: massive data scale, rapid data flow and dynamic data system, diverse data types and huge data value. Big data is a huge, high growth rate and diversified information asset that needs new processing mode to have stronger decision-making power, insight and process optimization ability. The strategic significance of big data technology lies not in mastering huge data information, but in specialized processing of these meaningful data. If big data is compared to an industry, the key to making profits in this industry is to improve the processing capacity of data.

Machine learning has become a new frontier subject and formed a course in colleges and universities. It integrates applied psychology, biology and neurophysiology as well as mathematics, automation and computer science to form the theoretical basis of machine learning. The unification of machine learning and AI is taking shape. For example, the idea that learning is combined with problem solving and knowledge expression is easy to learn produces the block learning of general intelligent system SOAR. Case-based learning, which combines analogical learning with problem solving, has become an important direction of empirical learning. Learning is a complex intelligent activity. The learning process and reasoning process are closely linked. According to the amount of reasoning used in learning, machine learning strategies can be generally divided into mechanical learning, imparting learning, analogy learning and case learning. The more reasoning used in learning, the stronger the ability of the system.

In many areas, such as the Internet and finance, the number of training instances is very large, and it is common to aggregate data sets of billions of events every day. In addition, more and more devices include sensors. Continuous recording of observed data can be used as training data. Such data sets can easily reach hundreds of TB. At present, the global and China's big data are showing blowout explosive growth. Big data has penetrated into various industries and business functions,

and become an important factor of production. The evolution of big data has a direct relationship with the improvement of productivity.

## II. LARGE DATA PROCESSING IN MACHINE LEARNING

Machine learning method is a method that computers use existing data to get a model and use it to predict. This method is very similar to human thinking. One of the main purposes of machine learning is to transform the process of human thinking and summing up experience into the process of computer processing and calculating the data to get the model. Generally speaking, machine learning algorithms can be divided into supervised learning, unsupervised learning, semi-supervised learning, Reinforcement learning and recommendation.

Big data will become an important strategic resource for all kinds of institutions and organizations, even at the national level. Attaching importance to the collection, mining, sharing and utilization of data resources has become a top priority. The opening and sharing of big data has become the trend of the times. Government departments must take the lead. The change of organization and global governance have become the inevitable choice.

The contribution of machine learning to the return on investment of large data applications is mainly reflected in two aspects: one is to promote the productivity of data scientists; the other is to find some neglected schemes, some of which are even ignored by the best data scientists. These values derive from the core function of machine learning: to enable analytical algorithms to learn the latest data without human intervention and explicit procedures. In many cases, machine learning is the best return on investment for big data innovation. Investing in machine learning can deepen any large data case customized to the enterprise. Deep learning has become an important tool in machine learning instruction system for big data scientists. In-depth learning using neural networks can help extract perception from these data streams, because these data streams may involve the hierarchical arrangement of the semantic relationships between constituent objects.

Automation is the key to understanding log data in depth, because log data is distributed on a large scale in the field of large data. Automation ensures data acquisition, analysis and processing. At the same time, it regulates data display results as fast as event-driven performance and data flow. The main engine of log analysis automation includes machine data integration middleware, business rule management system, semantic analysis, data flow computing platform and machine learning algorithm.

Different machine learning technologies are suitable for different types of log data and different analysis challenges. It is the correct way to construct a priori supervisory scheme for machine learning mechanism by using correlation and other existing modes. The unsupervised reinforcement learning mechanism may be more suitable if the log data schema cannot be precisely defined in a predictive way. These log data analysis schemes supported by machine learning technology are the most ideal scenarios for automated processing, because these schemes will independently select and prioritize highly matched processing modes, thus completing the established tasks without artificial training data sets.

## III. MACHINE LEARNING IN THE AGE OF BIG DATA

In the development of machine learning, there have been two major research directions. The first is to study the learning mechanism, focusing on exploring and simulating human learning mechanism; the second is to study how to effectively use information, focusing on acquiring hidden, effective and understandable knowledge from huge amounts of data. With the increasing demand for data analysis in the era of big data, the efficient acquisition of knowledge through machine learning has gradually become the main driving force for the development of machine learning technology.

Neural network model can form boundary for classification in the feature space, so as to build the nonlinear discriminant. When used in the classification problem, every feature vector element is

corresponding to one input node and the output node of each classification corresponding to is taken as the category of the distribution. Hidden nodes connect to input nodes, distributing initial weight and adjusting weight in the training process of neural networks. That multi-layer perception learning algorithm adopts the error back propagation method is based on the error function of the difference between the output and expecting output.

The weight among nodes in the neural network weight can be adjusted by error back propagation so as to reduce the output error of the network. Input/output neurons control is based on the input/output vector. Hidden layer neurons can be adjusted on the base of network performance requirements to adjust. According to the observation, the main reason for the long time neural network training is the great amount of data in network training. But this kind of intrusion detection system can classify the attack types online.

The goal of genetic algorithm is to find the approximate solution of the optimization problem. Genetic algorithm uses the climbing method to select any genes number, and there are four operations: initialization, selection, crossover and mutation. A single chromosome contains genes corresponding attributes, such as service, mark, login, or super user intent. Genetic algorithm could produce classification rules, and at the same time choose a parameter optimization in the testing process.

Genetic algorithm includes the following steps in the application of network data: 1) the intrusion detection system collects the corresponding communication network information via specified network. 2) The intrusion detection system applies rule to learn genetic algorithm training. 3) The intrusion detection system uses rules to do abnormal and normal classification on the communication information based on the model.

Machine learning has become a supporting technology and service technology. How to deeply analyze complex and diverse data based on machine learning and make more efficient use of information has become the main direction of current machine learning research. Machine learning is developing towards the direction of intelligent data analysis, and has become an important source of intelligent data analysis technology. In addition, in the era of big data, with the continuous acceleration of data generation, the volume of data has unprecedented growth, and new types of data that need to be analyzed are also emerging, such as text understanding, text emotion analysis, image retrieval and understanding, graphics and network data analysis, machine learning and research. Many new research directions have emerged in the research field, and many new machine learning methods have been proposed and widely used.

## VI. CONCLUSION

Big data refers not only to massive data, but also to unstructured, incomplete data that can not be processed by traditional methods. With the coming of the era of big data and the explosive growth of industrial data, the concept of big data has attracted more and more attention. With the explosive growth of industrial data, the concept of big data has attracted more and more attention. Research on machine learning algorithms in large data environment has become a topic of common concern in academia and industry. Machine learning is another important research field of artificial intelligence application after expert system, and it is also one of the core research topics of artificial intelligence and neural computing. With the explosive growth of industrial data and the unprecedented accumulation of data, the concept of Bigdata has attracted more and more attention. Big data is bringing huge profits to data-intensive enterprises. Big data is the guarantee of the upgrading of existing industries and the birth of new industries.

## REFERENCES

- I. Singh J, Nene M J. A Survey on Machine Learning Techniques for Intrusion Detection Systems. International Journal of Advanced Research in Computer and Communication Engineering, 2013(11):4349-4355.
- II. PAIVA R M, CARVALHO A R D, CRAWFORD C, et al. Comparison of surrogate models in a multidisciplinary optimization framework for wing design. American Institute of Aeronautics and Astronautics Journal, 2010, 48(5):995-1006.
- III. NAMJOSHI J, GUPTE A. Service Oriented Architecture for Cloud Based Travel Reservation Software as a Service. Proceedings of the 2012 IEEE International Conference on Cloud Computing, Sep 21-25, 2009, Bangalore, India. Los Alamitos, CA, USA: IEEE Computer Society, 2012:147-150
- IV. LAPLANTE P A, ZHANG J, VOAS J. Distinguishing Between SaaS and SOA. IT Professional, 2011, 10(3):46-50.
- V. Yao Liu. "Recovery compared with new technology", Posts & Telecom Press, 6, 2008.
- VI. Yuefeng Ji, Junjun Zou, Lin Bai, Daxiong Xu. "The experimental study of optical packet switching matrix", Acta Optica Sinica, 03, pp. 52, 2009.
- VII. Jinyi Xiong. "Packet switching technology and light burst switching technology", Scientific and Technological Information, S2, pp. 93, 2006.