The Application of Machine Learning in Data Mining

Zhenhang Wang, Xiaomeng Wei, Jilin Yang

School of Computer Science & Cyberspace Security, Hainan University, Haikou, China

Keywords: Data, Data mining, Machine learning, Technology applications.

Abstract: Machine learning is a multi-disciplinary subject that has emerged over the past 20 years and involves many disciplines such as probability theory, statistics, approximation theory, analytical theory, and computational complexity theory. Machine learning theory is primarily about designing and analyzing algorithms that allow computers to automatically "learn". Machine learning algorithms are a class of algorithms that automatically analyze and obtain rules from data and use rules to predict unknown data. Because learning algorithms involve a large number of statistical theories, machine learning is particularly closely related to inferred statistics, also known as statistical learning theory. In terms of algorithm design, machine learning theory focuses on achievable, effective learning algorithms. Many inference problems are difficult to follow without program, so part of the machine learning research is to develop an approximation algorithm that is easy to handle.

1. Introduction

Data Mining (DM) refers to the process of extracting useful information and knowledge hidden from a large number of incomplete, noisy, fuzzy, and random data. Data mining is one of the most cutting-edge directions in the field of database and information decision-making in the world. It has attracted extensive attention from academia and industry, and has been successfully applied in some fields.

Machine learning has been widely used in data mining, computer vision, natural language processing, biometrics, search engines, medical diagnostics, detection of credit card fraud, securities market analysis, DNA sequence sequencing, speech and handwriting recognition, strategy games and robotics.

2. Data mining and machine learning concept

2.1 Data mining.

It is a kind of decision support process and a kind of deep data analysis method. It is based on AI, machine learning, statistics and other technologies, highly automated analysis of the company's original data, inductive reasoning, mining potential patterns, predicting customer behavior, helping corporate decision-makers adjust market strategies and reduce Risk, make the right decisions. The business application of data mining can be described as: exploring and analyzing a large amount of enterprise data according to the established business objectives of the enterprise, revealing hidden, unknown or verifying the known regularity, and further modeling the advanced and effective methods [1]. Data mining is a process that uses various analysis tools to discover relationships between models and data in massive data. These models and relationships can be used to make predictions.

2.2 Overview of machine learning.

Anything that allows machines to learn about the relationships or rules between data through the models and algorithms we build, the techniques that we use at the end are machine learning techniques. In fact, machine learning technology is a cross-disciplinary subject, which can be roughly divided into two categories: traditional machine learning technology and deep learning technology, in which deep learning technology includes neural network related technologies. In this course, the focus is on traditional machine learning techniques and various algorithms [1].

  DOI: 10.25236/isrme.2019.074

Machine learning is a method of data analysis that automatically analyzes the building of a model, shown as Fig.1. By using an iterative learning data algorithm, machine learning allows a computer to discover hidden areas without being explicitly programmed to see.

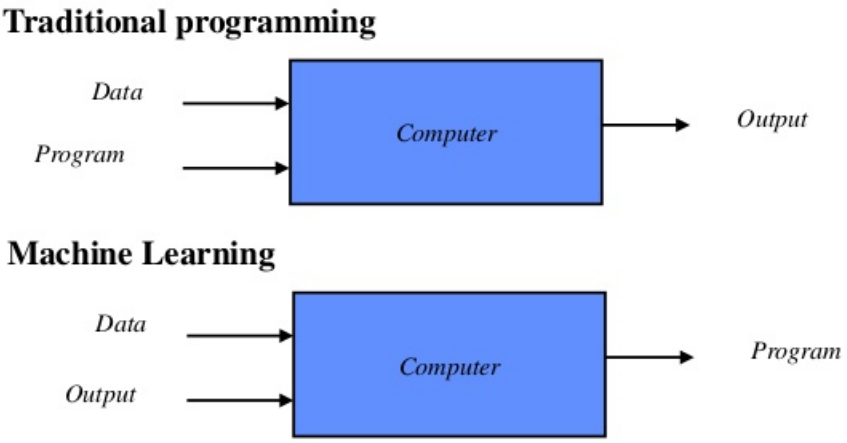

Fig.1 Characteristics of mechanical learning

Iteration is very important in machine learning. Because of its existence, the model can adapt to the data independently when it encounters new data. They can learn from previously generated reliable calculations, repeated decisions and results. Machine learning is not a completely new discipline - it is a discipline that gains new momentum.

Due to the emergence of new computing technologies, today's machine learning is very different from the past. Although many machine learning algorithms have been around for a long time, the ability to automatically apply complex mathematical calculations to big data (one after another, faster and faster) is the latest development.

Mechanical learning combines computing power with a special type of neural network to learn complex patterns in large amounts of data. Mechanical learning techniques currently work best in identifying objects in images and words in sound. Researchers are now looking for ways to identify these successful patterns to more complex tasks such as automated language translation, medical diagnosis, and many other important social and business issues.

3. Machine learning learning method

The two widely learned machine learning methods are supervised learning and unsupervised learning. Most machine learning (about 70%) is supervised learning. Unsupervised learning accounts for about 10%-20%. Semi-supervised and reinforcement learning techniques are sometimes used.

3.1 Supervise learning.

The algorithm uses the tag instance to train, just like the input that is known to be the desired output. For example, a device can have data points marked as “F” (failed) or “R” (running). The learning algorithm receives a series of inputs with corresponding correct outputs, and the algorithm learns by comparing the actual output with the correct output to find the error [2]. Through classification, regression, prediction, and gradient enhancement methods, supervised learning uses patterns to predict the value of additional unlabeled data labels. Supervised learning is commonly used to predict historical events that may occur in the future. For example, it can predict when credit card transactions can be fraudulent, or which insurance customer may file a claim.

3.2 Unsupervised learning.

Use the opposite data without a history tag. The system will not be notified of the “correct answer.” The algorithm must figure out what is being rendered. The goal is to explore the data and find some internal structures. Unsupervised learning works well for transactional data. For example, it can identify groups of customers with the same attributes (can be treated the same in marketing). Or it can find the main attributes to distinguish the customer base from each other. Popular techniques include

self-organizing maps, nearest-neighbor mapping, k-means clustering, and singular value decomposition [2]. These algorithms are also used for segment text topics, recommending items, and determining data outliers.

3.3 Semi-supervised learning.

The application is the same as supervised learning. But it uses both tagged and unlabeled data for training - usually a small amount of tagged data and a lot of unlabeled data (because unlabeled data is not expensive and can be obtained with less effort) . This type of learning can use methods such as classification, regression and prediction [3]. Semi-supervised learning is used when a fully tagged training process is too costly for the relevant tags.

3.4 Reinforce learning.

Often used for robotics, games and navigation. Through reinforcement learning, the algorithm discovers the maximum return from actions through trial and error. This type of learning has three main components: an agent (learner or decision maker), an environment (all agent interactions), and an action (what an agent can do). The goal is to act on behalf of the agent to maximize the expected reward in a given time [1]. With a good strategy, the agent will reach its goal faster. Therefore, the goal of reinforcement learning is to learn the best strategies.

4. The relationship between data mining and machine learning

The difference between machine learning and other statistics and learning methods, such as data mining, is another hot topic of debate. In simple terms, although machine learning uses many of the same algorithms and techniques as data mining, one of the differences lies in the predictions of these two disciplines:

Data mining is the discovery of previously unknown patterns and knowledge. Machine learning is used to reproduce known patterns and knowledge, automatically apply to other data, and then automatically apply these results to decisions and actions, as shown in Fig. 2.

The increasing power of computers is also stimulating the evolution of data mining for machine learning. For example, neural networks have been used for data mining applications for a long time. As computing power increases, you can create many layers of neural networks [4]. In machine learning languages, these are called “deep neural networks.” It is the improvement of computing power that ensures that automatic learning quickly handles many neural network layers.

Further, artificial neural networks (ANNs) are simply a set of algorithms based on our understanding of the brain. ANNs can - in theory - simulate any kind of relationship in a data set, but in practice it is very tricky to get reliable results from neural networks. The study of artificial intelligence dates back to the 1950s-labeled by the success and failure of neural networks.

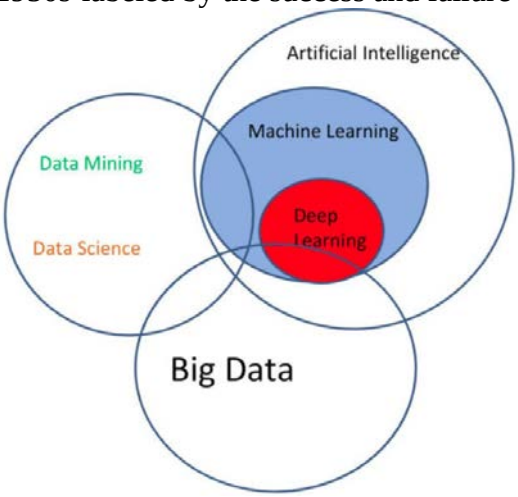

Fig.2 Characteristics of data mining and machine learning

5. The application of mechanical learning in data mining

Since machine learning technology and data mining technology are all exploring the laws between data, people usually put the two together. The application of mechanical learning in data mining is mainly reflected in the following points.

5.1 Dividing and verifying data.

Machine learning and data mining techniques can be used to solve classification problems, such as dividing customer levels, verification code identification, and automatic screening of fruit quality.

Taking verification code identification as an example, it is now necessary to design a scheme for identifying a verification code consisting of handwritten digits from 0 to 9. One solution is to first divide some of the handwritten numbers from 0 to 9 into training sets, and then manually divide the training set, that is, map each handwriting to its corresponding digital category, and establish these mappings. After that, the corresponding model can be established by the classification algorithm [4]. At this time, if a new digital handwriting appears, the model can predict the number represented by the handwriting, that is, which digital category it belongs to. For example, if the model predicts that a handwriting belongs to the category of number 1, the handwriting can be automatically recognized as a number 1. Therefore, the verification code identification problem is essentially a classification problem.

5.2 Data regression analysis prediction.

In addition to classification, data mining technology and machine learning technology have a very classic scene - regression. In the scenario of the classification mentioned above, the number of categories has certain limits. For example, the digital verification code recognition scene includes a numerical category of 0 to 9; and, for example, the letter verification code recognition scene includes a limited category of a to z. Whether it is a numeric category or a letter category, the number of categories is limited.

Now suppose that there is some data. After mapping it, the best result does not fall at a certain 0, 1 or 2 point, but continuously falls on 1.2, 1.3, 1.4... The classification algorithm can't solve this kind of problem, and then you can use the regression analysis algorithm to solve it [5]. In practical applications, the regression analysis algorithm can realize prediction and trend prediction for continuous data.

5.3 Data clustering.

As mentioned above, in order to solve the classification problem, it is necessary to have the category corresponding to the historical data, and the classification algorithm and the regression algorithm cannot solve the problem [5]. At this time, there is a solution - clustering, which directly divides the corresponding category according to the characteristics of the object. It does not need to be trained, so it is an unsupervised learning method.

When can I use clustering? If there is a group of customer's characteristic data in the database, it is now necessary to directly classify the customer's level according to the characteristics of these customers (such as SVIP customers, VIP customers), and then you can use the clustering model to solve [6]. In addition, when predicting the business circle, you can also use the clustering algorithm.

5.4 Data association analysis.

Correlation analysis refers to the analysis of the correlation between items. For example, if there is a large amount of goods stored in a supermarket, it is now necessary to analyze the correlation between the goods, such as the degree of correlation between the bread products and the milk products [6]. At this time, an association analysis algorithm can be used, with the help of users. Information such as purchase records directly analyzes the correlation between these products. After understanding the relevance of these products, they can be applied to the supermarket's merchandise placement, and by placing the highly correlated merchandise in a similar position, the merchandise sales of the supermarket can be effectively improved.

In addition, correlation analysis can also be used for personalized recommendation techniques. For example, by means of the user's browsing record, the association between the various web pages is analyzed, and when the user browses the webpage, the strongly associated webpage can be pushed to the user. For example, after analyzing the browsing record data, it is found that there is a strong relationship between the webpage A and the webpage C. When a user browses the webpage A, he can push the webpage C to him, thus realizing personalized recommendation [6].

6. Summary

Data mining technology can already be applied to all fields and industries. Data Mining Technology Data Mining Technology can be used in almost every aspect of people's lives. Not only does it bring great changes and influences to our daily lives, but it also profoundly changes our way of life. Although data mining technology has been applied to a certain extent and has achieved remarkable results, there are still many unresolved problems. With the deep research on mechanical learning, mechanical learning will surely be applied in a wider range of data mining technologies, and achieve more significant results.

References

[1] Y.G. Hu, Overview of Visualization Technology in Data Mining, Computer and Modernization, 2014, vol.4, pp.13-16.

[2] H.W. Lu, A review of the development and application of data mining, Journal of Changchun University, 2014, vol.10, pp.113~116.

[3] Q.T. Jiang, Application Analysis Based on Data Mining Technology, Changsha University of Science and Technology, 2016, vol.4, pp.102~104.

[4] Y.B. Zhang, Application of Data Mining Technology in Enterprise Decision-making, Journal of Information Science, 2018, vol.12, pp.42-45.

[5] Y.P. Yang, Data Mining Technology Research, Microcomputer Applications, 2017, vol.3, pp.65-68.

[6] W.T.Hao, Analysis of the application of data mining technology, microcomputer application, 2010, vol.9, pp.32-34.