From Statistics to Data Mining: A Brief Review

Jiachen Liu
School of Management
Beijing Sport University
Beijing, China
1751119@bsu.edu.cn

***Abstract*—With the popularity of the Internet and the development of information technology, hundreds of millions of information and data that are so close to our human kind are recorded. In recent decades, as an interdisciplinary, data science has had a vigorous development. This paper mainly discusses two core problems in data science, that are data analysis and data mining. Beginning with the data analysis based on statistics, this paper discusses and reviews data mining based on traditional machine learning and that based on neural network. Then it discusses the common problems in data mining. Each part summarizes the corresponding application and research. Through the discussion of the basic contents in data analysis and data mining, this paper provides a powerful help for the future research in this field.**

Keywords—data science, data mining, machine learning, review

I. Introduction

Basic statistics obtains the law of quantity by collecting, sorting and analyzing statistical data. Most statistical analysis is based on probability. However, with the development of computer technology, many problems are not merely about probability, but are the problems of incomplete information hidden. Data analysis based on statistics is not able to solve these problems, so new tools are needed to conduct the actual information computation, namely data mining. We use data mining to search for hidden information. The problem discussed in this paper is the development process from basic statistics to data mining. Related theories and algorithms are discussed in the following parts of this paper.

Data mining can dig out the essence of things, and it has been more widely used in all walks of life. For example, it can use algorithms to generate the spatial map of dust sources and use data mining models and integrated models to predict the storm dust sources [1]. The applications of data mining involve lots of fields, like customer and marketing, manufacturing, financial studies, etc. [2] In anti-doping, Bayesian framework is used to predict whether athletes use stimulants to improve the detection rate and comupting efficiency [3]. To sum up, the applications of data mining range widely, like medicine, education, sports, engineering and other fields, so it has a high research value.

The research value of this paper is to untangle the basic theories and related algorithms throughout the of data mining development, the application scenarios of data mining and the problems in previous studies, so as to provide reference for the use and future development of data mining. This paper will start with the introduction of relevant background research and discuss from the four aspects of data analysis based on statistics, data mining based on traditional learning methods, data mining based on neural network and the problems existing in data mining. And it summarizes and discusses the whole paper at last.

II. Data analysis based on statistics

Statistics is a method of analyzing data, which processes and refines the information in data. Statistics has two main categories, descriptive statistics and inferential statistics. Descriptive statistics is a statistical method of data collection, processing and description. For example, if you want to know the basic situation of a group of numerical data, you can describe its tendency of dispersion by calculating the average value, mode, and media, and complete the description with variance and standard deviation. In addition, the concept of distribution is introduced into descriptive statistics. Through the distribution map of data points, we can observe the skewness and peak value of the data, so as to calculate the probability of the data we need. Inferential statistics shows the inter or intra relationships among sample data by calculations. We can use point estimation and interval estimation to speculate through the known data, for example, by calculating the probability of an event in the sample data, we can speculate the probability that we cannot accurately measure or that of the event will occur in the future.

III. Data mining with traditional machine learning

Traditional machine learning provides approaches that can judge the significance of variables through regression and other methods but cannot predict. According to that, data mining provides the possibility of prediction through known variables and learnt parameters. Traditional machine learning consists of supervised learning models (classifying) and unsupervised learning models (clustering). The results of supervised learning are classified into classification and regression. The classification is conducted by extacting features from data sample (training set), and grouping targets (test set) according to the labbles or descriptions. Classical methods include k-Neares Neighbor (KNN), decision tree algorithm, Bayesian algorithm, case-based reasoning algorithm, and genetic algorithm, etc. Regression is a method which based on statistical principle for trend predictiion by establishing correaltion fuhctions between variables and attributes. It mainly includes unitary linear regression, the influence of one economic variable on another, and multiple linear regression, the correlation analysis of two or more independent variables and one dependent variable.

Unsupervised learning model is divided into clustering model and association rule. Clustering means grouping data into different sets according to their similarities nor sepecific features. The algorithm includes K-Means, K-Medoide and so on. Association rules are to mine hidden vaules or realtionships between data. The algorithm includes Apriori algorithm, FP-Growth algorithm and so on.

978-1-7281-7106-7/20/$31.00 ©2020 IEEE
DOI 10.1109/CDS49703.2020.00073

Data mining based on traditional learning methods can solve many problems, including the computer vision in IT field, such as face recognition, image search, natural language processing, and social network analysis which has user portrait, network association analysis, and recommendation system. Taking recommendation as an example, its algorithms can be based on popularity, demographic, content, and collaborative filtering. They are suitable for different scenarios, but all are based on the past records of users to predict the content that users may be interested in.

In addition, this method is widely used in other fields. For example, in the medical field, Marcin Czajkowski [4] proposed an advanced decision tree method, Evolutionary Multi-Test Tree (EMTTree), for discovery of biomarkers, the new gene-gene interaction and high-quality gene prediction. In Berrouiguet Sofian's paper [5], they stated that data mining can help with the detection of potential financial losses in medical insurance, thus, reducing the number of examine doctors and the workload of investigators.

IV. Data mining with neural network

With the development of the Internet, the data we are faced with becomes very complex. Now many data are not only simple numbers, but also images, audio, etc. Using the first two methods cannot achieve the effect we want, so we need to use neural network. The core of traditional machine learning is regression. Neural network is obviously different from tradition. Instead of simple regression, it learns different parameters and trains variable weights to build the model through repeated training.

Artificial Neural Network (ANN) is an artificial intelligence technology based on connectionism mechanism. ANN tries to simulate the micro cognitive function of human (neuron connections), to complete complex calculations. Based on sufficient training, expert experience and mathmatical principle, ANN acquires the possibility of pattern recognition. ANN overcomes the shortage that the decision-making interface is unable or inaccurate to count, which happens to the conventianal pattern recognition.

Because ANN has the functions of learning, association, self-organization, memory and fault tolerance, it can not only avoid the establishment of complex mathematical models and complicated mathematical reasoning, but also can run training and processing of ANN models for data with incomplete information, which can often obtain better results than conventional methods. As an important way and method to simulate human's intelligence and the ability of image thinking, ANN has achieved remarkable results in pattern recognition, signal processing, automatic control and other fields, and has broad application prospects in nonlinear modeling and nonlinear problem solving.

The basic structure of neural network consits of neurons and connections. Information is stored in the neurons in the form of features and weights. And connections decide the information passing. When the input is transferred into the network, extracted features will be saved and the network will generate the memory for the next stimulation.

Neural network is divided into feedforward network, feedback network and self-organizing network. At beginning, there is no feedback in the whole network, and the signal goes for a one-way communication from the input layer to the output layer. Feedforward network is the first generation ANN, based on the perceptron. The feedback network is based Hopfield's model, providing functions of associative memory and optimization calculation. Specific, ANN with back propagation (BP) is the second generation, which provides power abality in prediction and pattern recognition. Another, self-organizing neural network is a Self-Organizing Map (SOM) scheme which is based on the physiological laws of human brain. SOM network consists of input layer and competition layer. According to its learning rules, the network automatically classifies the input mode. That is, in the case of no teaching instruction, through repeated learning of the input mode, features can be contained in each input mode, Then the network organizes itself and displays the results in the competition layer. The difference between this representation and other types of networks is that it does not reflect the classification results with a neuron or a neuron vector, but with several neurons at the same time. This kind of network has strong anti-jamming ability and does not need training samples, so it is easy to realize. However, due to its high computational complexity and sensitivity to the similarity measurement, it cannot be applied to data sets with missing values.

At present, the problems solved by neural network include many aspects. For example, Aoki Genta mentioned in his article [6] that convolutional neural network (CNN) has been used for DNA sequence classification and proposed a new application of CNN in the classification of sequence arrangement in pairs to achieve accurate sequence clustering, which shows the advantages of CNN input in-pair arrangement for non-coding RNA (ncRNA) sequence clustering and module discovery. In Li Jingmei's paper [7], a screening algorithm of dangerous goods transportation is established by using Genetic Algorithm and Levenberg-Marquardt Neural Network (GA-LM-NN). For the transportation of dangerous goods in a single distribution center, a comprehensive optimaztion model is established to minimize the transportation risk and time, which has the ability to adjust its robustness. The algorithm can quickly find a set of transportation route for dangerous goods, and find Pareto optimal solutions of distribution routes with different robustness. Yu Ming [8] optimizes the identification performance of the existing algorithms for four kinds of electrocardiosignal, including sinus rhythm, ventricular fibrillation, ventricular tachycardia and cardiac arrest by using the back propagation neural network optimized by genetic algorithm. The accuracy of balance in the test set is as high as 99.06%. The application of this algorithm in the automatic external defibrillator will further improve the reliability of pre-defibrillation rhythm analysis, and ultimately improve the survival rate of cardiac arrest.

V. Common issues of data mining and solution

A. *Insufficient sample space and inaccurate data*

The lack of sample space and data is a common problem in data mining. It may even cause other problems such as over fitting. There are many ways to solve this problem, such as continuing to collect more data, reducing the data dimension to improve data generalization ability, or applying the method of data increment, using the synthetic "similar data" for training, such as image data that can be zoomed and flipped.

Walid Masoudimansour [9] proposed a new way of dimension reduction of the marked proportional data, which relieved the problem of data sparsity. Mengmeng Li [10] reformed a new method of fast hybrid dimension reduction, improving its speed, combining the function of multi-strategy feature selection with grouped feature extraction.

B. Computer problems

The workload of analyzing and processing a large number of data is huge, and its calculation efficiency and accuracy will affect the user's work efficiency and work judgment. When the scale of data increases and the dimension raises, the computing performance will drop dramatically. Now, most of the data belong to large-scale and high-dimensional data sets. The advanced distributed computing technology can be taken to solve this problem on the Hadoop distributed computing platform.

According to Jinhai Zhang [11], there was great redundance in massice data calculation with data warehouse. Then the continuous growing of the amount of data would increase the noise moreover. In this case, it was hard to conduct data mining via data warehouse, keeping promising efficiency and effectiveness. To overcome the massive computing and data storage, the Map/Reduce model was proposed. Fig. 2 shows an example of data mining with warehouse.

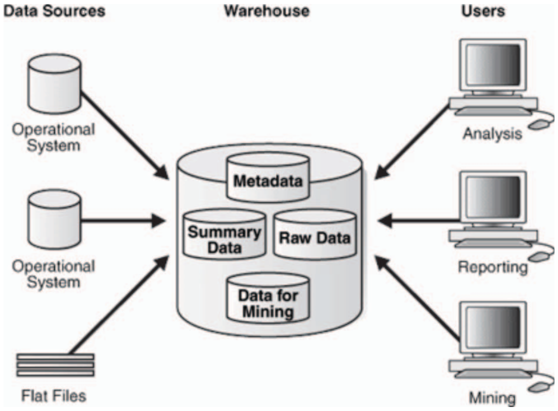

Fig. 1. An example of warehouse

C. Fitting problems

The overfitting problem is one of the most serious problems in machine learning. Its main consequence is that the low accuracy of invisible data models deteriorates the accuracy of data training. As long as the model is based on biased data, there will be overfitting problems. And when the model is complex and contains a large parameter set, it will lead to a more serious overfitting, such as the multi-layer neural network. In view of the overfitting problem in deep learning, one of the references [12] mentioned to solve the problem of converging to the maxima and overfitting of the Expectation Maximization (EM) algorithm by introducing an algorithm using nonparametric bootstrap program to enhance traditional EM [13]. Convolution neural network is used to detect workpiece defects. Aiming at the overfitting problem of small data sets in the practical application of deep learning, an iterative deep learning method was proposed to improve the recognition rate and reduce data overfitting [14]. A feature extraction algorithm based on multi-layer automatic encoder is proposed, which can prevent overfitting by the generative pre-training of data and regularization [15]. The influence of overfitting on data prediction can be effectively reduced by connecting to the dropout behind the connection layer of deep neural networks.

Underfitting happens when the lack of learning rate or the large amount of complex data. Generally speaking, underfitting problems can be solved by adding new features, reducing regularization parameters and using nonlinear models. Gary M. Bernstein [16] proposed in his paper a new measurement technology, which could reduce the measurement deviation and underfitting. In another reference [17], with the help of Ising model, an evaluation scheme was proposed for illustrating the efficiency of learning rate and regularization rate in recognition method. Thus, specific guidance for preventing overfitting and underfitting was provided when training and computing in ANN. The demonstation of underffting and overfitting is shown in the Fig. 2.

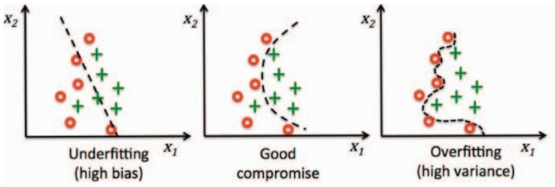

Fig. 2. Underffting and overfitting

VI. Conclusion

Presently, data mining has developed to wide application in different fields, and promoted the technical development in various fields. From the statistical perspective, this paper summarizes the development of data analysis and data mining, as well as their corresponding application fields. Besides, the paper summarizes the common problems in data mining, which are mainly from the perspective of methods, that is, the problem of machine learning and deep learning themselves. In addition, facing a widening application field, the research topics in the field of data science mainly focus on two aspects: data analysis and mining of complex information and distributed computing of information. Compared with the problems mentioned in this paper, these two aspects are more complex and comprehensive.

References

[1] Hamid Gholami, Aliakbar Mohamadifar, Adrian L. Collins. Spatial mapping of the provenance of storm dust: Application of data mining and ensemble modelling[J]. Atmospheric Research,2020,233.

[2] Xin Ma. Application of Data Mining in the Field of Human Resource Management: A Review[C].Proceedings of 1st International Symposium on Economic Development and Management Innovation(EDMI 2019),2019:236-241.

[3] Montagna Silvia, Hopker James. A Bayesian Approach for the Use of Athlete Performance Data Within Anti-doping.[J]. Frontiers in physiology,2018,9.

[4] Marcin Czajkowski, Marek Kretowski. Decision tree underfitting in mining of gene expression data. An evolutionary multi-test tree approach[J]. Expert Systems With Applications,2019,137.

[5] Berrouiguet Sofian, Ramírez David, Barrigón María Luisa,Moreno-Muñoz Pablo, Carmona Camacho Rodrigo, Baca-García Enrique,Artés-Rodríguez Antonio. Combining Continuous Smartphone Native Sensors Data Capture and Unsupervised Data Mining Techniques for Behavioral Changes Detection: A Case Series

of the Evidence-Based Behavior (eB2) Study.[J]. JMIR mHealth and uHealth,2018,6(12).

[6] Aoki Genta, Sakakibara Yasubumi. Convolutional neural networks for classification of alignments of non-coding RNA sequences.[J]. Bioinformatics (Oxford, England),2018,34(13).

[7] Li Jingmei, Wu Weifei, Xue Di, Gao Peng. Multi-Source Deep Transfer Neural Network Algorithm.[J]. Sensors (Basel, Switzerland),2019,19(18).Tengfei Li, Yongbin Qin. Feature Detection Base on Iterative Deep Learning [J].Computer and Digital Engineering,2017,45(06):1133-1137.

[8] Yu Ming, Chen Feng, Zhang Guang, Li Liangzhe,Wang Chunchen,Zhan Ningbo, Gu Biao, Wei Jing,Wu Taihu. [Research on malignant arrhythmia detection algorithm using neural network optimized by genetic algorithm].[J]. Journal of biomedical engineering,2017,34(3).

[9] Walid Masoudimansour, Nizar Bouguila. Supervised Dimensionality Reduction of Proportional Data Using Mixture Estimation[J]. Pattern Recognition,2020.

[10] Mengmeng Li, Haofeng Wang, Lifang Yang, You Liang, Zhigang Shang,Hong Wan. Fast hybrid dimensionality reduction method for classification based on feature selection and grouped feature extraction[J]. Expert Systems With Applications,2020,150.

[11] Jinhai Zhang. Design and Implementation of Data Mining Based on Distributed Computing. 2014, 3468:1702-1705.

[12] Marcin Czajkowski, Marek Kretowski. Decision tree underfitting in mining of gene expression data. An evolutionary multi-test tree approach[J]. Expert Systems With Applications,2019,137.

[13] Tengfei Li, Yongbin Qin. Feature Detection Base on Iterative Deep Learning [J].Computer and Digital Engineering,2017,45(06):1133-1137.

[14] Zhijun Sun, Lei Xue, Yangming Xu. Marginal Fisher Feature Extraction Algorithm Based on Deep Learning [J]. Journal of Electronics and Information Technology, 2013, 35(04): 805-811.

[15] ASHIQUZZAMANA, TUSHARAK, ISLAMMR, et al. Reduction of overfitting in diabetes prediction using dep learning neural network[C]//IT Convergence and Security, 2017:35-43.

[16] Gary M. Bernstein. Shape measurement biases from underfitting and ellipticity gradients[J]. Monthly Notices of the Royal Astronomical Society,2010,406(4).

[17] Andrei Dmitri Gavrilov, Alex Jordache, Maya Vasdani, et al. Preventing Model Overfitting and Underfitting in Convolutional Neural Networks. 2018, 10(4):19-28