Research on Statistical Problems in the Era of Big Data

Wenjia Deng

Ningbo Xiaoshi High School E Campus, Ningbo, Zhejiang, 315012, China

Catheringqtxs@163.com

Keywords: Big Data; Statistical Problems; Data Mining; Data Analysis; Machine Learning; Deep Learning; Statistical Methods

Abstract: With the advent of the big data era, statistics is facing unprecedented challenges and opportunities. This paper aims to discuss the statistical problems in the era of big data and propose solutions. First, we outline the definition, characteristics and impact of big data on statistics, and point out the challenges posed by big data to traditional statistical methods. Then, we classify and analyze the big data statistics problems, including data acquisition and preprocessing, data quality and accuracy, data analysis and modelling, data visualization and interpretation, etc. Then, we discussed the methods to solve the problem of big data statistics, including the application and limitations of traditional statistical methods, the application of machine learning and deep learning in big data statistics, and the impact of big data technology on statistics. Finally, we summarize the research, point out the existing problems and deficiencies, and put forward the future development direction and suggestions. This paper aims to provide a reference for understanding statistical problems in the era of big data and guidance for research and practice in related fields.

1. Introduction

1.1. Research Background

In today's digital and information age, big data has become one of the core driving forces in various fields. With the rapid development of information technologies such as the Internet, the Internet of Things and social media, a large amount of data is continuously generated, stored and transmitted. These data are not only large in scale but also of various types, covering all aspects of human social life, such as economy, medical treatment, education, environment, etc. However, owning big data alone is far from enough. How to extract valuable information from massive data and make correct decisions has become an important task before us.

In this context, statistics, as a discipline that studies data collection, analysis, interpretation and inference, has become particularly important. Traditional statistical methods can no longer meet the need for efficient analysis of big data. Therefore, statistical problems in the era of big data are becoming more and more prominent. Traditional statistical methods may face problems such as slow computing speed and insufficient model complexity when processing big data, and it is difficult to mine the rules and trends hidden behind the data. Therefore, it has become an urgent problem to study how to use statistical methods to solve practical problems in the era of big data [1].

1.2. Research Significance

The significance of this study lies in:

First, it fills the blank of statistical research in the era of big data. With the rapid development of big data, traditional statistical methods are no longer competent for the task of analyzing and explaining massive data. Therefore, it is of great theoretical and practical significance to explore statistical methods suitable for big data.

Second, it will promote the cross-integration of statistics and other disciplines. The application of big data has been involved in various fields, including finance, healthcare, and e-commerce. A large amount of data needs to be analyzed and mined in these fields. Therefore, the cross-integration of statistics and these fields will promote the generation of new research results.

Finally, it is of great significance to improve the efficiency and accuracy of data analysis. By studying the statistical problems in the era of big data, more efficient and accurate data analysis

DOI: 10.25236/icamfss.2024.019

methods can be explored, thus providing a more reliable basis for decision-making.

1.3. Research Purposes

The purpose of this study is:

First, it discusses the statistical problems in the era of big data. By analyzing the characteristics of big data and the limitations of statistical methods, we can identify the statistical problems in the era of big data and develop solutions. [2]

Second, it proposes methods to solve the problem of big data statistics. Based on the analysis of the characteristics of big data and the limitations of statistical methods, statistical methods and techniques suitable for big data are proposed to solve practical problems.

Finally, it explores empirical research on big data statistics. Through the analysis of actual cases in various fields, the effectiveness and feasibility of the proposed solution are verified, guiding practice in related fields.

2. Overview of Statistical Issues in the Era of Big Data

2.1. Definition and Features of Big Data

Big data refers to a collection of data that requires new data processing methods due to its large data volume, complexity, diversity, and fast update. Compared with traditional data processing methods, big data has the following main features:

First, big data is huge scale. The "large" here refers not only to the size of the data volume but also to the size that traditional data processing methods cannot process. The size of big data is usually measured at TB, PB, EB, and even larger ZB and YB levels.

Second, big data has various types. Big data includes not only structured data but also semi-structured and unstructured data. Structured data refers to data that can be organized and stored in a certain format, such as table data in a database; Semi-structured data refers to data that has a certain structure but does not conform to the traditional database table format, such as XML, JSON, and other formats; Unstructured data refers to data that does not have a clear structure, such as text, image, audio, video, and so on.

Third, big data is fast generation. With the rapid development of information technologies such as the Internet, the Internet of Things and social media, a large amount of data is continuously generated and updated. For example, social media platforms generate a large amount of user comments, posts, pictures, videos and other data every day; Internet of Things devices continuously generate various sensor data; Financial transaction markets generate a large amount of transaction data every day.

Fourth, big data has a low-value density. Most of the data in big data may be useless, including a large amount of noise data and invalid information. This requires extracting valuable information and knowledge from data analysis and mining.

Finally, there are processing and Analysis challenges for big data. Due to the large scale, diverse types, and fast generation of big data, traditional data processing and analysis methods are no longer competent. Therefore, how to efficiently process and analyze big data has become an important challenge.

2.2. Big Data Challenges to Statistics

The big data era has brought many challenges to statistics, mainly in the following aspects:

First, data acquisition and preprocessing is an important issue. The acquisition of big data involves data collection, storage, cleaning and other processes, including the transmission and storage of massive data, as well as data cleaning and de-noising and other pre-processing work. Due to the large scale and various types of big data, data acquisition and preprocessing become complicated.

Secondly, data quality and accuracy are key issues in big data analysis. Big data often has low data quality problems, such as data missing, noise, outliers, etc. These problems may have a great impact on the results of data analysis and modelling. How to ensure the quality and accuracy of big data has become an urgent problem to be solved.

In addition, data analysis and modelling are also important challenges. Traditional statistical methods may face problems such as slow computing speed and insufficient model complexity when processing big data, and it is difficult to mine the rules and trends hidden behind the data from massive data. How to develop efficient and accurate data analysis and modelling methods suitable for big data is a problem that needs in-depth study. [3]

Finally, data visualization and interpretation are also important challenges. In the era of big data, more intuitive and efficient data visualization methods are needed to quickly discover valuable information from massive data and explain and explain the results. Traditional data visualization methods are often unable to meet the needs of the big data era, so it is necessary to develop new data visualization methods suitable for big data.

2.3. The Development Process of Big Data Statistics Methods

With the advent of the big data era, statistical methods are also continuously developing and evolving, mainly reflected in the following aspects:

First, there are improvements in traditional statistical methods. According to the characteristics and challenges of big data, researchers have improved and optimized traditional statistical methods to make them more suitable for big data analysis and modelling. For example, to solve the problem of high-dimensional features of big data, researchers have proposed a series of dimensionality reduction methods and feature selection methods to reduce the complexity of the model and the computational time cost.

Second, the application of machine learning and deep learning is advancing. As a new data analysis method, machine learning and deep learning have been widely used in the era of big data. Their powerful data processing and pattern recognition capabilities provide new ideas and methods for big data analysis. For instance, deep learning has achieved remarkable results in fields such as image recognition and natural language processing, offering new possibilities for analyzing and applying big data.

In addition, the impact of big data technology on statistics is also an important aspect. The rapid development of big data technology has also had a profound impact on statistics, providing more powerful and efficient tools and platforms for statistics. For example, the development of distributed computing, parallel computing, and storage technologies enables statisticians to process and analyze larger and more complex data sets. The progress of these technologies has provided important support and guarantee for the development of statistics.

In addition, the exploration of emerging statistical methods and tools is also an important direction for the development of statistics in the era of big data. With the advent of the big data era, researchers are constantly exploring new statistical methods and tools to adapt to the changing big data environment. For example, the graph model-based statistical method can effectively process large-scale network data. The stream data analysis method can process and analyze data streams in real-time and is suitable for real-time data monitoring and decision-making [4].

3. Classification and Analysis of Major Data Statistics Problems

3.1. Data Acquisition and Preprocessing

In the era of big data, data acquisition and preprocessing are one of the primary challenges faced by statistics. Big data acquisition involves collecting massive amounts of data from multiple sources and integrating them into data warehouses that can be analyzed. However, access to big data often faces many difficulties and challenges. First of all, the data sources of big data are diversified, which may come from sensors, social media, mobile devices and other channels. How to efficiently collect and integrate these data has become a complex problem. Secondly, the data volume of big data is huge, and traditional data transmission and storage methods may not meet the requirements. Therefore, distributed storage and processing technologies are needed. In addition, the data formats of big data may be various, such as structured data, semi-structured data and unstructured data. How to effectively process these different types of data is also an important issue.

The preprocessing phase is one of the key steps in big data analysis. It includes data cleansing, data integration, data conversion, and data protocol. Data cleansing refers to clearing abnormal values such as errors, duplicates, and deletions in data to ensure data quality and accuracy. Data integration refers to the integration of data from different sources into a unified data set for subsequent analysis and modelling. Data conversion refers to the conversion of data from one form to another to meet different analysis requirements. Data specification is to reduce data complexity and improve data processing efficiency by reducing dimension and clustering. [5]

3.2. Data Quality and Accuracy

Data quality and accuracy are one of the key issues in big data analysis. Big data often has low data quality problems, such as data missing, noise, outliers, etc. These problems may have a great impact on the results of data analysis and modelling. The quality of data directly affects the accuracy and reliability of data analysis and modelling results. Therefore, how to ensure the data quality and accuracy of big data has become an urgent problem to be solved.

Data quality problems may come from many aspects, such as errors in data collection, losses in data storage, and interference in data transmission. Among them, the error in the process of data collection is a common problem, which may come from the error of sensors, the failure of data collection equipment, the error of human operation, etc. Loss during data storage refers to data loss or damage caused by hardware failure, software failure, and other reasons. Interference in data transmission refers to errors in data transmission due to network delay, network congestion, and other reasons. [6]

To ensure the data quality and accuracy of big data, a series of measures can be taken. First of all, it is necessary to carry out quality inspection and correction on the data to find and correct errors and outliers in the data in time. Secondly, a perfect data management system needs to be established, including data collection, storage, transmission and processing. Thirdly, efficient data processing and analysis methods are needed to improve data processing efficiency and accuracy. Finally, it is necessary to strengthen data security and privacy protection to protect data security and privacy from infringement.

3.3. Data Analysis and Modeling

Data analysis and modelling are one of the core links in big data analysis. In the era of big data, data analysis and modelling often face various challenges and difficulties. First of all, traditional data analysis methods may not be competent due to the large scale, diverse types and fast generation of big data. Traditional statistical methods may face problems such as slow computing speed and insufficient model complexity when processing big data, and it is difficult to mine the rules and trends hidden behind the data from massive data. [7] Secondly, data analysis of big data often needs to be implemented with the help of high-performance computing and distributed computing technologies, which puts forward higher requirements on computing resources and algorithm efficiency. Thirdly, data analysis of big data often needs to be realized by combining domain knowledge and professional technology, which puts forward higher requirements for the ability and quality of data analysts.

When facing the problems of data analysis and modelling big data, a series of methods and technologies can be adopted. First, machine learning and deep learning can be used to analyze and model data. As an emerging data analysis method, machine learning and deep learning have powerful data processing and pattern recognition capabilities, which can help us mine the rules and trends hidden behind the data from massive data. Secondly, high-performance computing and distributed computing technologies can be used to analyze and model big data. High-performance computing and distributed computing technologies can improve the efficiency and speed of data processing and analysis, thus speeding up the process of data analysis and modelling. Thirdly, the data analysis and modelling process can be guided by domain knowledge and expertise. Domain knowledge and expertise can help us better understand the meaning and background of data to conduct data analysis and modelling more accurately. Finally, data can be analyzed and displayed with the help of visualization technology. Data visualization is an intuitive and intuitive data analysis method, which can help us quickly discover the rules and trends in data and visually display the analysis results.

3.4. Data Visualization and Interpretation

Data visualization and interpretation is one of the important links in big data analysis. In the era of big data, data visualization can not only help us better understand data, but also help us discover the rules and trends in data and visually display the analysis results. However, data visualization and interpretation face many challenges and difficulties. First of all, traditional data visualization methods may not be competent due to the large scale, diverse types and fast generation of big data. Traditional data visualization methods are often limited to processing small data sets, but cannot process large amounts of big data. Secondly, data visualization of big data often requires high-performance computing and distributed computing technologies, which require high computing resources and algorithm efficiency. Thirdly, data visualization of big data often needs to be realized by combining domain knowledge and professional technology, which puts forward higher requirements for the ability and quality of data visualization personnel. [8]

When facing the problem of data visualization and interpretation of big data, a series of methods and technologies can be adopted. First, you can use visualization software and tools to visualize and interpret data. Visualization software and tools can help us quickly visualize data and visually display the analysis results. Secondly, high-performance computing and distributed computing technologies can be used to visualize and interpret big data. High-performance computing and distributed computing technologies can improve the efficiency and speed of data processing and analysis, thus accelerating the process of data visualization and interpretation. Thirdly, we can use domain knowledge and professional technology to guide the data visualization and interpretation process. Domain knowledge and expertise can help us better understand the meaning and background of data to visualize and interpret data more accurately. Finally, interactive visualization technology can be used to visualize and interpret data. Interactive visualization technology allows users to participate in data analysis and display to better understand data and analysis results.

4. Solutions to Major Data Statistics Problems

4.1. Application and Limitations of Traditional Statistical Methods

Traditional statistical methods still have certain application value in the era of big data, but they also face some limitations. Traditional statistical methods include various probability statistical models, hypothesis testing methods, regression analysis, etc. They have good applicability and interpretability when analyzing small-scale data or when the data volume is relatively small. However,in the face of big data, the limitations of traditional statistical methods are gradually emerging.

First of all, traditional statistical methods often need to process and analyze the entire data set, while the data volume of big data is huge, and traditional methods may not be competent. For example,traditional regression analysis methods often consume a lot of computing resources and time when facing big data, and even cannot complete the analysis. Secondly, traditional statistical methods often assume that data have certain distribution patterns and correlations, but data in big data often have high complexity and heterogeneity, which may not be satisfied. Thirdly, traditional statistical methods are often unable to process nonlinear and high-dimensional data, while data in big data often have nonlinear relations and high-dimensional characteristics, which limits the application of traditional methods.

Nevertheless, traditional statistical methods still have certain advantages and values. For example, in the case of analyzing small-scale data or requiring causal inference, traditional statistical methods are still an effective tool. Therefore, the combination of traditional statistical methods and emerging data analysis methods can better solve the problem of large data statistics.

4.2. The Application of Machine Learning and Deep Learning in Big Data Statistics

As a new data analysis method, machine learning and deep learning have important application value in big data statistics. Machine learning and deep learning can efficiently process and analyze big data by learning the patterns and rules between data, thus solving some limitations of traditional statistical methods when facing big data.

First, machine learning and deep learning have strong data processing and pattern recognition capabilities. They can process large amounts of big data and dig out hidden rules and trends. For example, deep learning has achieved remarkable results in fields such as image recognition and speech recognition, providing new possibilities for the analysis and application of big data.

Secondly, machine learning and deep learning can effectively process nonlinear and high-dimensional data and are suitable for complex and diverse data types in big data. For example, machine learning methods such as support vector machines and decision trees can handle nonlinear relationships and high-dimensional features and are suitable for data analysis and modelling in big data.

Thirdly, machine learning and deep learning have better generalization and adaptive capabilities, can adapt to different types and scales of data, and produce better prediction and classification effects. For example, a neural network model in deep learning can automatically adjust model parameters according to data characteristics to adapt to different data distributions and data characteristics.

In general, machine learning and deep learning have significant application value in big data statistics, which can help us better process and analyze big data, and extract valuable information and knowledge from it.

4.3. Impact of Big Data Technology on Statistics

The rapid development of big data technology has had a profound impact on statistics, providing new opportunities and challenges for the development of statistics. First, big data technology provides statistics with more powerful and efficient tools and platforms. For example, the development of distributed computing, parallel computing, storage technology and other technologies enables statisticians to process and analyze larger and more complex data sets. Secondly, big data technology has changed the research methods and ideas of statistics. Traditional statistical research is often based on small sample data and sample surveys, while big data technology enables statisticians to use full sample data for research and analysis, thus obtaining more accurate and comprehensive conclusions. Thirdly, big data technology has promoted the cross-integration of statistics and other disciplines. The development of big data technology enables statisticians to better cooperate and communicate with computers, information science, artificial intelligence and other fields to jointly solve practical problems and promote the development and innovation of disciplines. For example, the cross-research of statistics and computer science has produced new research fields such as data mining and machine learning, providing new ideas and methods for the analysis and application of big data.

In addition, big data technology has also had an impact on statistics education. Traditional statistics education often pays attention to the explanation of theoretical knowledge and mathematical methods but ignores the practical ability of data processing and analysis. However, in the era of big data, statisticians need to have strong data processing and analysis capabilities and be able to skillfully use various data analysis tools and technologies for practical work. Therefore, statistics education needs to keep pace with the times and pay attention to cultivating students' data analysis and practical ability so that they can adapt to the needs of the big data era.

4.4. Exploration of Emerging Statistical Methods and Tools

With the advent of the big data era, researchers are constantly exploring new statistical methods and tools to adapt to the changing big data environment. The exploration of emerging statistical methods and tools covers various fields, including data mining, machine learning, deep learning, text analysis, image analysis, etc. These emerging methods and tools not only expand the research field of statistics but also provide new ideas and methods for the analysis and application of big data.

First of all, data mining is a method to discover hidden patterns and knowledge in data, which has been widely used in business, medical treatment, finance and other fields. Data mining methods include classification, clustering, and association rule mining, which can help us mine valuable information and knowledge from big data.

Secondly, text analysis is a method to analyze and understand text data, which has been widely used in natural language processing, public opinion analysis, emotion analysis and other fields. Text Analysis methods include text classification, text clustering, emotion analysis, etc., which can help us extract valuable information and knowledge from a large amount of text data.

Thirdly, image analysis is a method to analyze and understand image data, which has been widely used in image recognition, image segmentation, target detection and other fields. Image analysis methods include feature extraction, feature matching, depth learning, etc., which can help us identify and analyze valuable information and knowledge from image data.

In general, the exploration of emerging statistical methods and tools provides new ideas and methods for the analysis and application of big data and injects new vitality and motivation into the development of statistics. Through continuous exploration and innovation, we are confident that we can give full play to the role of statistics in the era of big data and make greater contributions to the development of data science and human society.

5. Conclusion and Prospect

5.1. Research Summary

The advent of the big data era has brought unprecedented opportunities and challenges to statistics. In this paper, the statistical problems in the era of big data are deeply studied and discussed, and the characteristics of big data, challenges to statistics, and the development process of statistical methods are discussed in detail. Through the classification and analysis of big data statistics problems and the discussion of solutions, this paper has important theoretical and practical significance for understanding and dealing with statistical problems in the big data era.

First of all, we have made an in-depth analysis of the characteristics of big data and discussed the characteristics of large scale, various types and fast generation of big data. These characteristics determine the difference between big data and traditional data processing methods, laying a foundation for us to understand the statistical problems in the era of big data.

Secondly, we conducted a comprehensive analysis of the statistical challenges posed by big data, including data acquisition and preprocessing, data quality and accuracy, data analysis and modelling, data visualization and interpretation, etc. These challenges are important problems faced by statistics in the era of big data, which require us to seriously consider and explore solutions.

Then, we delved into the methods of solving big data statistics issues, which include the application and limitations of traditional statistical methods, the application of machine learning and deep learning in big data statistics, the impact of big data technology on statistics, as well as the exploration of emerging statistical methods and tools. These methods provide important references and enlightenment for us to deal with statistical problems in the era of big data.

5.2. Existing Problems and Deficiencies

However, we are also aware of some problems and deficiencies in this paper. First of all, this paper is more in-depth in theoretical analysis and method discussion but lacks empirical research and case analysis. Empirical research and case analysis can more intuitively show the actual situation and solutions to statistical problems in the era of big data, thus making the paper more convincing and operable. Secondly, there may be some situations that are not clear and perfect enough in the elaboration and demonstration of some contents in this paper, which need to be further refined and improved. Finally, the discussion on the future development direction and suggestions in this paper is not deep and specific enough and needs to be further expanded and deepened.

5.3. Future Development Direction and Suggestions

Because of the above problems and deficiencies, we put forward the following future development directions and suggestions:

First, it is suggested to strengthen empirical research and case analysis. Through in-depth analysis and research on the actual cases of big data statistics problems, the essence and solutions of the problems can be displayed more intuitively, thus improving the persuasiveness and operability of the paper.

Secondly, it is necessary to further refine and perfect the paper's content. The theoretical analysis and method discussion of the paper need to be further refined and improved to make the logic of the paper clearer and the demonstration more sufficient.

Thirdly, it is necessary to further discuss the future development direction and suggestions. In the discussion of future development direction and suggestions, it is necessary to analyze the development trend and challenges of big data statistics in more depth and put forward more specific and operable suggestions and measures to point out the direction for the development of statistics in the era of big data.

In general, the big data era has brought new opportunities and challenges to statistics. We need to continuously explore and innovate, strengthen theoretical research and practical exploration, and make greater contributions to the development of statistics in the big data era.

6. Conclusion

In the era of big data, statistics is facing unprecedented challenges and opportunities. This article systematically outlines the definition, characteristics, and profound impact of big data on statistics, and then analyzes in detail the key issues encountered in the process of big data statistics, such as data collection and preprocessing, data quality and accuracy, data analysis and modelling, and data visualization and interpretation. In response to these challenges, this article delves into the limitations of traditional statistical methods and discusses in detail the application of advanced technologies such as machine learning and deep learning in big data statistics, as well as the innovative driving role of big data technology in statistics.

When summarizing research results, we emphasize that to address the challenges of the big data era, it is necessary to further strengthen empirical research, improve theoretical systems, and deepen interdisciplinary integration and innovation.

We hope that these suggestions can provide guidance for research and practice in related fields, help statistics better respond to the challenges of the big data era, and seize the opportunities it brings. We firmly believe that through continuous exploration and innovation, statistics will demonstrate broader application prospects and profound academic value in the era of big data.

References

[1] Halevy, A., Norvig, P., & Pereira, F. (2009). The unreasonable effectiveness of data. IEEE Intelligent Systems, 24 (2), 8-12.

[2] Chen, M., Mao, S., & Liu, Y. (2014). Big data: A survey. Mobile Networks and Applications, 19 (2), 171-209.

[3] Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer Science & Business Media.

[4] Jordan, M. I ., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. Science, 349(6245), 255-260.

[5] Wu, X., Zhu, X., Wu, G. Q., & Ding, W. (2014). Data mining with big data. IEEE Transactions on Knowledge and Data Engineering, 26 (1), 97-107.

[6] Goodfellow, I ., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press.

[7] Kitchin, R. (2014). Big data, new epistemologies and paradigm shifts. Big Data & Society, 1 (1), 2053951714528481.

[8] VanderPlas, J. T. (2016). Python Data Science Handbook: Essential Tools for Working with Data. O 'Reilly Media.