Improved Statistical Analysis Method Based on Big Data Technology

Hongsheng Xu[1,2*]
[1]Luoyang Normal University
[2]Henan key Laboratory for Big Data Processing & Analytics of Electronic Commerce
Henan LuoYang, China
E-mail: 85660190@qq.com

Ke Li
Luoyang Normal University
Henan LuoYang, China
E-mail: 85660190@qq.com

Ganglong Fan[1,2]
[1]Luoyang Normal University
[2]Henan key Laboratory for Big Data Processing & Analytics of Electronic Commerce
Henan LuoYang, China
E-mail: 85660190@qq.com

***Abstract*—Big data technology refers to the rapid acquisition of valuable information from various types of large amounts of data. It can be divided into 8 technologies: data acquisition, data access, infrastructure, data processing, statistical analysis, data mining, model prediction and results presentation. The paper presents improved statistical analysis method based on big data technology. A statistical analysis model in big data environment is designed to extract useful information features from large amounts of data based on the Hadoop system by using its distributed storage and parallel processing mechanism.**

Keywords-Big data; Statistical analysis; Hadoop; Data acquisition; Data mining

I. INTRODUCTION

With the rapid growth of the scale of statistical data, data characteristics become increasingly complex, data collection channels are diverse, and statistical related field research has entered the era of big data. How to efficiently collect sample data, mine information, extract useful information features from large amounts of data, and provide information to relevant departments in a timely manner has become one of the focuses of current statistical research [1]. Compared with foreign countries, there are some problems in the statistical analysis of our country, such as the low degree of integration of information resources, the lack of data sharing and incomplete information. With the advent of the era of big data, research and application of data analysis and mining of large pay more and more attention, big data mining and analysis will help the statistics department in reasonable time collection, management and analysis of massive data.

Statistical analysis of the eight methods, one index, comparative analysis, index comparative analysis, also known as comparative analysis, is the most commonly used method of statistical analysis. It is a comparative method to reflect the differences and changes in the number of things.

Big Data refers to the large data size exceeds the commonly used software tools at run time can withstand the collection, management and data processing ability of data sets; data is currently stored mode and ability, computing and storage and processing capacity can not meet the existing data sets generated by the relative concept of scale.

With the development of information technology, more and more data are accumulated. In fact, the data itself is meaningless and can only really work if it is used for analysis. Therefore, it can be said that the more important behind the surge of data is implicit information, and people want to be able to analyze these data at a higher level in order to make better use of these data. The massive data is the development trend of data analysis and data mining is becoming more and more important, from the mass of data to extract useful information is important and urgent, this will require the processing to be accurate, high precision, and the processing time is shorter, get valuable information quickly, therefore, promising research of massive data, too worthy of extensive research.

In the large data environment, facing the collection and statistics of massive data, traditional methods can not meet the needs of large-scale data set processing [2]. Based on the Hadoop system, using the distributed storage and parallel processing mechanism, and it is the design of the statistical data environment analysis model, to extract useful features information from massive data, realizing the sharing of data resources, to provide information service for the relevant decision-making departments.

Statistics is a data processing engineering, dealing with large data sets, the statistical sample becomes large, complex data feature makes statistical work has become cumbersome, and data mining is a process to get useful information from a large number of data, the use of modern information technology and mining algorithm, can effectively useful for data acquisition and processing. It might be accurate data statistical model for processing data for a large data mining under the condition of understanding, relevant data processing and analysis of mining data after introducing the statistics, two kinds of methods are combined. The paper

presents improved statistical analysis method based on big data technology.

II. DISCUSSION ON THE INTEGRATION OF BIG DATA AND STATISTICAL ANALYSIS

Partial least squares regression is a new multivariate statistical analysis method. It was first proposed by Wood and Abano in 1983. In the past ten years, it has developed rapidly in theory, method and application. Many statisticians are beginning to focus on their theoretical research, and its growing potential in applications is attracting more and more attention.

According to the data contained in the prior information in the background, the data set can be divided into homogeneous and heterogeneous (homogeneity) (heterogeneity), this paper introduces the two kinds of data integration punishment analysis method; it summarizes both considering network structure (Network), the method of punishment. The regression coefficient of integration analysis has two meanings: the first is the variable level, and the ordinary single data set model; second is the data set level, the same explanatory variables with a regression coefficient associated with each data set is connected by the regression coefficient. This is also the particularity of integration analysis. The significance of variables is no longer a regression coefficient, but a set of regression coefficients. Therefore, a special variable selection method is needed.

Packet analysis comparison index contrast, but the overall statistical units have a variety of characteristics, which makes the unit in the same overall range have many differences, statistical analysis not only on the total number of features and quantitative analysis of the relationship, but also the overall was analyzed deeply inside. Packet analysis is based on the statistical analysis of the objective requirements, the overall research in accordance with one or several marks is divided into several parts, collate, observation, analysis, to reveal the inherent relationship between it.

Big data technology refers to the rapid acquisition of valuable information from various types of large amounts of data. It can be divided into 8 technologies: data acquisition, data access, infrastructure, data processing, statistical analysis, data mining, model prediction and results presentation. At the same time, three computing models, batch processing, stream processing and interactive analysis, are formed by these techniques, as is shown by equation(1) [3].

$$P(\beta;\lambda,\gamma)=\lambda\sum_{j=1}^{p}\left\|\beta_j\right\|^{\gamma}=\lambda\sum_{j=1}^{p}\left(\left(\sum_{i=1}^{M}(\beta_j^{(i)})^2\right)^{1/2}\right)^{\gamma} \quad (1)$$

The classification is to find out a set of data objects in the database of the common characteristics and in accordance with the classification model can be divided into different classes; its purpose is through the classification model, mapping the data item to touch a given category. Can be applied to the prediction of application involves classification, trends, such as the Taobao shops will be users over a period of time in the purchase are divided into different classes, recommended Association class products to users according to the situation, so as to increase the sales of shops.

Hadoop has developed into a collection containing multiple sub items. The core content is the MapReduce and Hadoop distributed file systems (DHFS). It also includes Common, Avro, Chukwa, Hive, Hbase, and other sub projects, they provide high-level services on the core layer, and play an important role in the promotion of Hadoop applications.

Big data has been defined as the fourth paradigm of scientific inquiry. After hundreds of years of experimental science, thousands of years ago before the theory of science and decades ago computational science, the data explosion gave birth to data intensive science, theoretical, experimental and computational simulation paradigm of unity. Big data has been hailed as "non competitive" factors of production. Big data has "inexhaustible," the characteristics of the continuous re-use, restructuring and expansion of the continuous release of its potential value, in a wide range of open, sharing, and constantly create new wealth. The root is that the value of big data is to predict future trends in unknown areas and non specific factors, to solve long-term, universal social problems. The current big data technology and applications are still limited to historical and real-time data association analysis, limited to meet short-term, specific market demand. The process of solving paradoxes is just the course of theory and method. While people try to solve the paradox of effort, just big data push air plant.

Data mining is from a large, incomplete, noisy, fuzzy and random data to extract implicit, believable, novel, people do not know in advance, but is potentially useful patterns of advanced treatment process. Data mining is a cross subject formed by the integration of many fields, such as statistics, artificial intelligence, database and visualization technology [4]. In addition to describing relationships and rules, one of the most important tasks of data mining is analysis. According to the laws found in past and present data, this model can sometimes be considered as a key attribute of time.

The partial least squares regression theory is the greatest contribution of Umea University Organic Chemistry Department wood its founder professor Wood taught the moral education in sweden. Under his guidance, as is shown by equation (2), and the Department has published many doctoral dissertations on the theory and applications of partial least squares regression. He and his collaborators have also conducted extensive theoretical discussions and developed SLMCA-P data analysis software running under Windows to support partial least squares regression calculations and interpretation of results. Perhaps this is true. Partial least squares regression is widely used in the field of chemical engineering.

$$\xi_{ij}(k)=\left[1+\left|\frac{\Delta x_i(k)}{\sigma_i}-\frac{\Delta x_j(k)}{\sigma_j}\right|\right]^{-1} \tag{2}$$

Where x is often in the use of ordinary multiple flyback when Wei, is the number of samples should not be too small. In the general statistics book, the number should be more than two times the number of variables. However, in some of the scientific research experiment, there are often many important variables must be considered, but because of the condition fee and time limit, the number of available samples is far less than the variable. The general multiple regression model is incapable of modeling when the number of sample points is less than the number of variables.

Integration analysis is also an effective way to solve the "size" problem. It integrates multiple data sets and increases the sample size. It is an effective way to solve the small sample problem. This problem is very common in big data, on the one hand, due to large data sparsely, low value density, the marginal value of information is not the amount of data with increased; on the other hand is the high dimensionality of the data highlight the Internet and cloud computing for data acquisition and storage to bring convenience, small study on the factors associated with the phenomenon may be collected, dimensions will be high, as is shown by equation(3), and "noise purification" is an urgent problem to be solved. Integrated analysis is a variable selection method combined with integration analysis, dimension reduction is an effective way to extract information, not only can be applied to model selection, correlation analysis between data sets can, in order to better identification of signal and noise.

$$P(\beta;\lambda,a,b)=\sum_{j=1}^{p}P_{MCP}\left(\sum_{m=1}^{M}P_{MCP}(\left|\beta_j^{(m)}\right|;\lambda,a);\lambda,b\right) \tag{3}$$

The key to processing and analyzing large data lies in the distributed storage function and powerful computing power. The basis of data processing is data storage, and the key to data analysis lies in the powerful processing ability. Hadoop is a scalable and reliable computing system, open source distributed, the framework can be realized by simple calculation model of massive data processing in computer cluster, compared with on high performance servers, Hadoop good scalability, while the nodes in the cluster can provide local storage and calculation.

III. IMPROVED STATISTICAL ANALYSIS METHOD BASED ON BIG DATA TECHNOLOGY

Time series is a series of values that change and develop in the same time in time. They are formed in chronological order, forming a time series, also called a dynamic series. It can reflect the development and change of social economic phenomena. Through the compiling and analysis of time series, we can find out the law of dynamic change, and provide the basis for predicting the future development trend [5]. The time series can be divided into absolute number, time series, relative number, time series, and average time series.

A large collection of data received from the client is using multiple databases (Web, App or sensor form) data, and the user can perform simple queries and processing work through these databases. For example, the electricity supplier will use traditional relational databases such as MySQL and Oracle to store every transaction data. In addition, NoSQL databases such as Redis and MongoDB are also commonly used for data collection. In the process of collecting data, the main characteristics and challenges is the high number of concurrent, because at the same time there may be tens of thousands of users to access and operate, such as train ticketing website and Taobao, visit their concurrent at the peak reached millions, so in the end need to support the deployment of a large number of data acquisition. And how to load and distribute between these databases requires deep thinking and design.

The missing value ratio, which is based on data columns that contain too many missing values, is less likely to contain useful information. Therefore, you can remove columns with missing data columns greater than a certain threshold. The higher the threshold is, the more efficient the dimensionality reduction method is, the less the lower dimension, the lower variance is similar to the filtering method, which assumes that the data column changes very little and the information contained in the column is very small. As a result, all columns with small variance are removed. One thing to note is that the variance is related to the range of data, so you need to normalize the data before using the method.

Data collection center is mainly through the deployment in the cloud server cluster environment to complete data acquisition, data are stored in HDFS distributed database; statistics management department to set up the server cluster, in order to ensure the scalability of the system, can also be incorporated into the base layer of the server at any time in the cluster computing tasks by using MapReduce the mechanism of distribution and processing; statistical analysis center is mainly intelligent algorithm pool, through the analysis of the application of algorithm for the data collection.

Statistics is an ancient discipline, has more than 300 years of history, in the natural science and social science development has played an important role in statistics; it is a strong vitality and discipline, as is shown by equation (4), and it all rivers run into sea with the growing development and learn widely from others'strong points, specific discipline each door. Without exception, the arrival of the big data era has brought opportunities for the development of statistical disciplines, but also made statistical disciplines face major challenges. How to deeply understand and grasp the development opportunity, how to better understand and deal with this great challenge, so we need to clarify the concept of "big data features clear big data; put forward the

new concept of statistical thinking process to re-examine the statistics [6].

$$w_{i+1}^{1}(t+1)=\left(1-wd_{i}^{1}(t)\right)x_{i}^{1}(t)-rs_{i}\alpha N^{1}(t) \quad (4)$$

The regression analysis reflects the attribute value of the data in the database, and finds the dependence between the attribute values through the relation between the function and the data mapping [7]. It can be applied to the prediction of the data sequence and the study of the correlation. In marketing, regression analysis can be applied to every aspect. Through the regression analysis of the quarterly sales, we forecast the sales trend in the next quarter and make targeted marketing changes.

Time series speed index. According to the absolute number of time series can be calculated speed indicators: there is development speed, growth rate, average speed of development, the average growth rate. Dynamic analysis is method. In statistical analysis, it is difficult to make a judgement if there is only one period index value. If the time series is worked out, dynamic analysis can be carried out to reflect the changing law of its development level and speed.

Hadoop provides a stable and reliable analysis system and shared storage for statistical analysis. It contains two core technologies: MapReduce and HDFS [8]. MapReduce implements data processing and analysis, and HDFS is responsible for data sharing and storage. In large data environment, the basic framework of statistical work includes data acquisition center and statistical analysis processing center.

IV. Experiments and Analysis

Analysis of statistics based on large data, the statistical object is often structured and unstructured mixed data, such as text, image, audio and video, here is the basic idea of the design is the use of the underlying mining model through data collection, management middleware, implementation layer analysis, screening and sorting out the valuable data and information finally, the statistical results of visualization.

Partial least squares regression can solve many problems that can not be solved by ordinary multiple regression. In the application of ordinary multiple linear regression, we are often faced with many restrictions, and the most typical problem is the multiple correlation between the self changing and the most [9]. Many experienced system analysts have noticed this problem. In order to describe and analyze systems more fully, as far as possible without omitting some of the most important system characteristics, analysts tend to selects relevant indicators more carefully.

Factor analysis is using the index. Factor analysis is the research object is divided into various factors, the overall research object as the factors common result, through the analysis of various factors, the influence degree of the research object in the general changes of factors were determined [10]. The factor analysis can be divided into the factor analysis of the change of the total index according to the statistical index of the object under study, and the factor analysis of the change of the average index.

Statistics and analysis of the main use of the distributed database, or distributed computing analysis and classification of common summary of mass data storage within the cluster, in order to meet the demand analysis of the most common, in this regard, some real-time requirements will be used EMC GreenPlum, Oracle Exadata, and MySQL based storage Infobright so, some of the batch, or based on semi-structured data needs can use Hadoop, as is shown by equation(5).

$$q_{ii}=\lim_{h\to 0^{+}}\frac{p_{ij}(h)}{h}=\begin{cases}\lambda_{i},\ \text{j=i+1},\\ u_{i},\ \text{j=i-1},\\ 0,\ |i-j|\geq 0.\end{cases} \quad (5)$$

The purpose of using large statistics is to infer the average or quantile of economic, social or social, economic or social indicators. The emphasis of statistics is on the representativeness of samples, which are generally met by probability sampling. Although there is a large data sample mass, can provide a wealth of information, but strictly speaking, big data is not a sample, on the contrary there will be lack of large data sample representation, information redundancy, noise and other problems, this situation is very easy to bring system error analysis results.

Big data based on the analysis of massive data to produce value, then how to get massive data to make big data really landing it. One of the most important aspects of this is data openness. Now to promote data openness and it is more importantly, through the sharing of data to produce more value. Data opening can improve the efficiency of social operation, and actively integrate the public data of all parties, and establish urban planning based on big data to ease traffic and social security issues. The opening of data can stimulate great commercial value, and the opening of data is open to the public, and anyone can use it to create new business opportunities.

People familiar with multivariate statistical analysis know that there are two broad categories of multivariate statistical analysis methods. One is the model based approach, which is mainly represented by regression analysis and discriminate analysis. It is characterized by the separation of independent variable and dependent variable in the set of variables. Data analysis is often used to find the functional relationship between dependent variables and independent variables. A model is established for prediction. The other is the cognitive method, which is represented by principal component analysis and cluster analysis, and canonical correlation analysis belongs to this method. The main feature of this kind of method is not in the original case according to the independent and dependent variables of the points, and through data analysis, can simplify the data structure and the similarity between observed variables or sample points.

V. SUMMARY

The paper presents improved statistical analysis method based on big data technology. Balance analysis is a method to study the equivalence of quantitative changes in social economic phenomena. It arranges the two sides of the unity of opposites according to their constituent elements, and gives the whole concept, so as to facilitate the whole situation to observe the balance relation between them. Balance relationship exists widely in economic life, to the national macro economy, small personal income. Balance analysis functions: one is the balance to reflect the social economic phenomenon from the number of equivalence relations, analysis of the ratio between the various phase to adapt to the situation; the two is to reveal the factors and development potential is not balanced; the three is the balance between the individual indicators can be calculated from the given index in the unknown.

At present the government with the e-government platform can realize the sharing of data resources, but the enterprise between the government and the lack of data sharing platform, causing the information isolation, in this regard, the statistics department to build a full range of safety statistics data sharing and distributed storage analysis platform, implementation of statistical information exchange across the region, and to meet the real time share mass data processing.

ACKNOWLEDGMENT

This paper is supported by Henan key Laboratory for Big Data Processing & Analytics of Electronic Commerce, and also supported by the science and technology research major project of Henan province Education Department (13B520155, 17B520026).

REFERENCES

[1] Patricia L. Mabry. Making Sense of the Data Explosion. American Journal of Preventive Medicine, 2011, 40(5),pp.12-30.

[2] Viktor Mayer-Schonberger,Kenneth Cukier,Big Data: A Revolution That Will Transform How We Live, Work and Think, Hodder & Stoughton,2013.

[3] Letouzey, S. Huberlant, P. Mares et al.. Assessment of Quality of Life of Patients Supported for Genital Prolapse Surgery: Feasibility of a Computerized Data Collection. The Journal of Minimally Invasive Gynecology, 2011, 18(6).

[4] W. Aigner, A. Rind, S. Hoffmann. Comparative Evaluation of an Interactive Time-Series Visualization that Combines Quantitative Data with Qualitative Abstractions,Computer Graphics Forum, 2012, 31,pp.3-15.

[5] B. Zhu, L. Xu, D. Faries et al.. PMH83 Comparison of Total Health Care Costs Between Remitters and Non-Remitters for Schizophrenia Patients from a Prospective Longitudinal, Observational Study in the Presence of Missing Data. Value in Health, 2012, 15(4),pp.100-120.

[6] Hassibi,Khosrow & De, Big Data, Data Mining, and Machine Learn, John Wiley Sons, 2014.

[7] Ahmed M. Abdel-Khalek, Mostafa A. Elseifi, Kevin Gaspard et al.. Model to Estimate Pavement Structural Number at Network Level with Rolling Wheel Deflectometer Data. Transportation Research Record: Journal of the Transportation Research Board, 2012, 2,pp.30-41.

[8] Lee, Keon Myung & Park, Seung Jong & Lee, Soft Computing in Big Data Processing, Springer, 2014.

[9] Yanqing Lv, Jianmin Gao, Zhiyong Gao and Hongquan Jiang, "Multifractal information fusion based condition diagnosis for process complex", Process Mechanical Engineering, (2012),pp.1-8.

[10] Bauckhage C, Kersting K. Data mining and pattern recognition in agriculture, KI-Künstliche Intelligenz, 2013, 27(4): 313-324