Проблеми статистичних обчислень, пов'язані з обсягом і багатомірністю даних, їх різноманітністю і високою швидкістю, та вимагають нових поглядів на парадигми статистичних і обчислювальних методів. Проаналізовано основні проблеми, пов'язані з використанням традиційних статистичних методів для аналізу багатовимірних даних, і показані області їх застосування. Розглянуто останні методичні досягнення в галузі статистики, що стосуються аналізу великих даних

Ключові слова: великі дані, статистичний аналіз, методи статистичного аналізу, бутстрап, передискретизації

Проблемы статистических вычислений, сопряженные с объемом и многомерностью данных, их разнообразием и высокой скоростью, требуют новых взглядов на парадигмы статистических и вычислительных методов. Проанализированы основные проблемы, связанные с использованием традиционных статистических методов для анализа многомерных данных, и показаны области их применения. Рассмотрены последние методические достижения в области статистики, касающиеся анализа больших данных

Ключевые слова: большие данные, статистический анализ, методы статистического анализа, бутстрап, передискретизации

UDC 004
DOI: 10.15587/1729-4061.2015.51603

REVIEW OF STATISTICAL ANALYSIS METHODS OF HIGH-DIMENSIONAL DATA

M. S. Hajirahimova
PhD*
E-mail: makrufa@science.az
A. S. Aliyeva
Researcher*
E-mail: aybeniz63@rambler.ru
*Institut Information Technology of ANAS
Vahabzade str., 9 B, Baku,
Republic of Azerbaijan, AZ1141

1. Introduction

Wide spreading of modern information technologies to all spheres of society, the expansion of people's access to the Internet, as well as increasing the number of services provided through the Internet network, overall wide application of the digital technology in as health, astronomy, bioinformatics, transportation, public administration and etc. cause for increases of data flow, including formation of "big data" phenomenon. This new term is used for to identify data that cannot be processed with in terms of volume and complexity of the existing management methods and by means of intelligent analysis [1]. Big Data's storage, management, value creation made a serious problem. The problem is automatically and continuously generated from different sources of data processing in real-time and irrationality of existing IT (Information Technology) solutions. The current situation allows us to express "to create data is extremely easy and to process data is extremely difficult" [1]. Current situation, application fields, opportunites, problems and etc. of big data technologies have been studied widely in [1–4].

Big data promise new level of scientific inventions and economic value. Thus, scientific achievements based on more and more to the management of data, researchers become consumers more. İn generally, big data techniques include a number of disciplines: statistics, data mining, machine learning, social network analysis, optimization methods, visualization approaches and etc. Statistical methods are very important for analyse of big data. Presented article is excatly dedicated to interpretation of these methods.

On the other hand for volume, intensity and complexity as well as big data exceed the capabilities of standard software, create problems for analysis of data with traditional statistical methods. It cannot be possible analyze of this kind of big data by current statistical methods in an ordinary computer. In this case, appear problems such as settle of large-scale big data's to an ordinary computer memory, deceleration of speed of the process, require more time. It is necessary to turn to the multi-core and cloud computing technology parallel and distributed architectures. At the same time, current capabilities of parallel and distributed architecture of data storage and management require statistical methods compliance with paradigm of big data. Also, by increase size of data, the complexity of their structure makes it necessary to apply new methods and puts great demands in front of the current statistical methodology.

2. Analysis of published data and problem statement

Big Data bring new chances to contemporary society and problems to data scientists. One of the main characteristics of big data is that the statistical methods, which work well on small-scale datasets, usually implement poorly in big data settings. So, it is not easy to hope for the performance of those statistical methods for the big data problems.

Large volume and high speed may bring heterogeneity, noise accumulation, counterfeit relations and random endogeneity, creating issues in computational feasibility and algorithmic stability [5]. Either with newly developed statistical methodologies and/or computational methodologies can be approached to these obstacles.

High variety brings nontraditional or even unstructured data types, which calls for new, creative ways to re-

alize the structure of data and even to ask intelligent study questions [6].

Big data current issues much further beyond the area of classic statistics, requiring joint workforce with domain knowledge, computing skills, and statistical thinking [7].

There are several algorithms that are lately improved and suitable for statistical inference of big data and workable on parallel machines, including the bag of little bootstraps [8], aggregated estimation equation [9], split-and-conquer algorithms [8, 10], and the subsampling-based stochastic approximation algorithm [11]. Moreover, recurrent algorithms have been widely used in the present society of scientific computing.

Samples of such iterative algorithms include different Markov chain Monte Carlo (MCMC) algorithms [12, 13], and the EM algorithm [9, 10], which typically require a large number of iterations and a complete scan of the full dataset for each iteration.

However, standard statistical techniques are normally not well fit to control Big Data and many researchers have offered expansions of classical techniques or absolutely new methods [14-18]. Authors proposed efficient approximate algorithm for large-scale multivariate monotonic regression [16], parallel statistics and several parallel statistics algorithms [15, 17].

But, in fact this kind of big data sets cannot be analyzed on a single commodity computer because their sizes are too big to fit in memory or it is too time consuming to process when the present statistical methods are used. To get over this barrier, one may have to apply to parallel and distributed architectures, with multicore and cloud computing platforms providing access to hundreds or thousands of processors. With thriving size usually comes a growing complexity of data structures, of the patterns in the data, and of the models needed to account for the patterns too. Big data has put a big challenge on the modern statistical methodology [10].

3. Purpose and objectives of the study

The key purpose of this paper is to analyze statistical methods for big data which cannot be analyzed and converted into a major scientific direction with help of available tools.

In accordance with the set goal the following research objectives are identified:

1. The application areas of big data are reviewed.

2. The statistical analysis problems are investigated in these fields.

3. The last methodological approaches are analyzed comparatively in solving the problems.

4. The application fields of big data

Big Data's transformative potentials are given in 5 domains by McKinsey institute in a report recently: healthcare; public sector administration; retail; global manufacturing and personal location data [2].

In the following subsections, we will shortly acquaint some applications of the Big Data problems in commerce and business, healthcare, economy and finance, and scientific research fields.

Big Data in scientific research. Many scientific fields have already become highly data-driven with the development of computer sciences. For example, astronomy, meteorology, social computing, bioinformatics and computational biology are based large volume of data with different types created or produced in these science fields [2].

The advent of big data has given rise to new research paradigm. In 2007 Turing Award winner Jim Gray, depicted a fourth paradigm of scientific research data volumes. He believed that the fourth paradigm can only be way to address some of the most difficult global challenges we face today [19].

Physicists learn the features of particles by colliding them with other particles in high-tech experiments. Other Big Data applications lies in numerous scientific subjects such as astronomy, atmospheric science, medicine, genomics, biologic, biogeochemistry and other complex and interdisciplinary scientific researches [2].

CERN is host to one of the largest familiar experiments in the world, as well as an example of big data, supremely. For more than 50 years, CERN has been struggling the increasing flows of data produced by its experiments researching main particles and the forces by which they interact. The Large Hadron Collider (LHC) consists of a 27-kilometer ring of superconducting magnets with a number of expediting structures to enhance the energy of the particles along the way. The detector sports 150 million sensors and acts as a 3D camera, taking pictures of particle collision events at the speed of 40 million times per second [20] and can generate 60 terabytes of data per day. The patterns in those data can give us an unprecedented understanding the nature of the universe. 32 petabytes of climate observations and simulations were conserved on the discovery supercomputing cluster in the NASA Center for Climate Simulation (NCCS) [2].

Acknowledging that this data probably keeps many answers to the mysteries of the universe that are being searched for a long period of time, and responding to the need to store, distribute and analyze the up to 30 petabytes of data produced each year, the Worldwide LHC Computing Grid ensure the requisite global distributed network of computer centers [19].

For instances, a sophisticated telescope is regarded as a very large digital camera which generate huge number of universal images. For example, the Large Synoptic Survey Telescope (LSST) will record 30 trillion bytes of image data in a single day. The size of the data equals to two entire Sloan Digital Sky Surveys daily. Astronomers will utilize computing facilities and advanced analysis methods to this data to investigate the origins of the universe [2].

In genetics, for instance, DNA gene sequencing machines based on big data analytics can now read about 26 billion characters of the human genetic code in seconds [21].

Authors in [2] noted that as various types of data are generated and produced in the field of Science and research. One common point exists in these disciplines is that they generate enormous data sets that automated analysis is highly required.

Big data in healthcare. Data is critical in the healthcare industry where it documents the history and evolution of a patient's illness and care, giving healthcare providers the tools they need to make informed treatment decisions. With medical image archives rises by 20 to 40 per cent per year. McKinsey analysts forebode that, if large sets of medical data were regularly collected and electronic health records

were filled with high-resolution X-ray images, mammograms, 3D MRIs, 3D CT scans, etc., we could better predict and provision to the healthcare needs of a population; which would not only drive gains in efficiency and quality, but also cut the costs of healthcare dramatically [20].

Analyzing large datasets of patient characteristics, results of treatments and their cost can help define the most clinically effective and cost-efficient treatments to use.

Research of big data in the field of genomics currently allows us to open genetic symptoms of rare diseases and link between the diseases and to find rare variants of consistency.

The field of neuroimaging has also witnessed big growth made in integratively analyzing imaging data of multiple subjects and multiple modalities, together with genomic data [8]. However, noisy images of the brain, the analyse of scanned data space hundreds of times of head movements creates great difficulties for the statistics and neurologists.

As in the area of geonomics to remove the systematic biases which caused by experimental variations and data aggregations is one of the main challenges.

Research of microbiom plays an important role in human health [5]. Next generation consistency technologies have made it feasible to learn all microbes in human intestine in an impartial method. Analysis of such large volumes of reads data, usually in 100s of terabytes, raises many challenges in statistical analysis and computation [5].

In addition, statistically controlled inclusion of a subject in a group study, i.e. testing if a person should be refused as outlier data, is often poorly conducted and voxels cannot be perfectly aligned across varied experiments in various laboratories [5, 6].

Consequently, authors in [5, 6] showed that the collected data contain a lot of outliers and missing values. These problems make data preprocessing and analysis substantially more compound. Many traditional statistical processes are not well suited in this boisterous high dimensional parameters, and mainly new statistical thinking is necessary.

Big data in economy and finance. Appropriately to the report from McKinsey institute [2] the efficient use of Big Data has the essential benefits to transform economies, and conveying a new wave of productive growth. Ever more corporations are taking the data-driven approach for conduct more targeted services, minimize risks and raise performance over the last ten years. They are performing specialized data analytics programs to collect, store, manage and analyze big datasets from a range of sources to determine key business insights that can be used to support better decision making.

It is difficult analyzing a large panel of economic and financial data. For instance, as an significant tool in analyzing the common evolution of macroeconomics time series, the customary vector autoregressive (VAR) model generally includes no more than 10 variables, given the fact that the number of parameters increases quadratically with the size of the model [5].

But, currently econometricians need to analyze multivariate time series with more than hundreds of variables. Including all information into the VAR model will cause hard over fitting and bad prognosis performance.

To resort to sparsity likelihoods, under which new statistical tools have been developed is one solution [5].

Another sample is folder optimization and risk management [5]. In this issue, evaluating the covariance and reverse covariance matrices of the returns of the assets in the portfolio plays an considerable role.

Authors in [5] noted that the cumulated error of the whole matrix estimation can be large under matrix norms even if we could estimate each individual parameter accurately and this demands new statistical procedures.

Big Data in commerce and business. Purchase-transaction data from commercial websites have long been gathered. But now new kinds of big data are generated by commercial websites. According to evaluates, the volume of business data worldwide, among companies, nearly doubles every 1.2 years [2]. This deluge of data can be discovered by using Big data, which help for example to optimize business processes, detect a pattern, better understand customers, anticipate their behaviors, needs and intentions.

Gradually, businesses base their decisions on data. Businesses need workers to collect convenient product data and analyse that data in the context of the industry. For deciding what kinds of improvements or new products they should make to meet their customers' needs analysts look at purchase data and customer reviews. For instance, to seeing what types of products customers buy and when they buy them workers may study transaction data from store loyalty cards. Big data can also help businesses run more efficiently. Analysts apply supply chain data to control inventories. They find faults by investigating real time production data too [2].

Analyzing and mining big data can also effectively safeguard public security and combat criminal and economic crimes, telecommunications and society administration. collection.

5. Problems of analysis of big data

The latest achievements in the field of modern digital technology caused widespread of large-scale data collection. Climate, social networking data, smartphones and data about health status, unstructured text data, social media and financial time series, e-commerce data, retail contract notes, surveillance videos including such data. The statistical analyses of such large-scale data collection have decisive importance.

Big data have unique features which traditional data doesn't have. Big data are characterized by large volume, high size of samples and highly growth rate. As a rule, are collected from several sources by various technologies at certain times. These feature resulting problems such as collection of noise, wrong correlations, computational complexity and instability of algorithms. It creates heterogeneous problems and leads to mistakes in experimental and statistical analysis [5, 8]. These features create significant problems for analysis of data and increasing of statistical methods. At the same time it makes difficult applying of traditional statistical procedures. For example applying of many traditional methods which are suitable for medium-sized data collection to big data is impossible. At the same time possible statistical methods for lower size data face important challenges during analysis of high-dimensional data.

So increasing volume and size of data, it is impossible keeping of them in a counting machine and effective processing by traditional statistical analysis methods.

Thus, during the statistical analysis of big data can include followings to encountered main problems [22]:

– large volume and diversity of big data;

– cannot applying for large-scale terabytes of data collection of a lot of popular algorithms of statistical analysis or very low processing speed;

– different structure of data;
– data protection and privacy issues;
– lack of time;
– not have any information extraction practices of statisticians.

Thereby, it does not make sense to expect the productivity of traditional statistical methods for the problems of big data. New statistical and computational methods are important for to solve the problems of big data.

6. Statistical methods for big data

2013 is The International Year of Statistics. It's an assignment for the purpose of highlight the role that data and statistical analysis have in society [23]. Statistics had achieved significant success in the research field of big data, too. There are several algorithms that are recently developed and feasible for statistical inference of big data and workable on parallel machines, including the bag of little bootstraps, aggregated estimation equation, split-and-conquer algorithms, and the subsampling-based stochastic approximation algorithm [6].

The aggregated estimation equation and split-and-conquer algorithms are based on the same idea of divide-and-conquer, but focus on different types of problems; the former is for parameter estimation and the latter for variable selection of regression models.

The recent methodologies for big data can be grouped into three categories [8]: resampling-based, divide and conquer, and sequential updating.

a. Subsampling-Based Methods

Subsampling based method which proposed for the statistical analysis of big data incorporates 3 approaches [8]:

1) Bags of Little Bootstrap (BLB);
2) Mean Log-likelihood;
3) Leveraging.

1) Bags of Little Bootstrap (BLB)

The bags of little bootstrap (BLB) approach have been offered by Kleyner and et al. for the statistical analysis of big data. Through this approach is possible to provide discrete assessments such as dispersion or confidence intervals and quality indicators. It is a combination of subsampling, the m-out-of-n bootstrap and the bootstrap to achieve computational efficiency [8, 24].

BLB consists of the following step [8]. Firstly, m-dimensional s – subsamples are taken from n-dimensional initial data. Discrete assessment and quality indicators (such as confidence intervals) are obtained from n-dimension r bootstrap samples taken for each of this s subsample. Then general discrete assessments and quality indicators are gotten from combination of s bootstrap point estimates (calculations) and quality measures (for example, by average).

In summary, BLB algorithm consists of 2 procedures which included to each other:

1) inner procedure is applied for initial loading (bootstrap) of subsample;

2) outer procedure incorporated many assessments of bootstrap (initial loading).

Though, the inner bootstrap procedure conceptually generates multiple resampled data of size n, what is really needed in the storage and computation is a sample of size m with a weight vector. Unlike the subsampling and the m-out-of-n bootstrap, in this approach there is no need to expand of the scale of confidence intervals until the final result for analytic correlation.

BLB procedure facilitates distributed calculation by gives an opportunity to processed of m-dimensional each subsample in separate processors. Kleyner and his colleagues proved BLB's dynamism and high level correctness. Their modeling studies showed high precision, convergence speed and the remarkable computational efficiency [8].

The bag of little bootstraps modifies the ordinary bootstrap to make it suitable for large-scale data sets. Though, the problem of processing massive bootstrap samples is facilitated, but computation of the estimates for a large number of bootstrap samples is excessively expensive in BLB [25].

BLB is unpractical for generally high complexity modern estimators. On the other hand, using the primitive LS estimator in the original BLB scheme does not provide a statistically robust bootstrap procedure [25].

Fast and Robust Bootstrap (FRB) method that proposed by Shahab Basiri et al. is scalable and compatible with distributed computing architectures and storage systems, robust to outliers and consistently provides accurate results in a much faster rate than the original BLB method [25].

2) Mean Log-likelihood

Liang et al. offered stochastic approximation approach that based on repeat selection with application of geostatistical data. The method uses Monte Carlo averages calculated from subsamples to approximate the quantities needed for the full data [13, 16].

The solution to the mean score equation is obtained from a stochastic approximation procedure, where at each iteration, the current estimate is updated based on a subsample of size m drawn from the full data. As m is much smaller than n, the method is scalable to big data.

Liang et al. created the sequence and asymptotic normality of the resulting estimator under soft situation. In a simulation research, the convergence speed of the method was nearly independent of n, the sample size of the full data [8, 11].

Liang & Kim stretched the average log-likelihood into a bootstrap Metropolis–Hastings algorithm in Markov chain Monte Carlo (MCMC) [12].

İn the Metropolis–Hastings algorithm the probability ratio of the suggestion and current estimate is rotated with that approximated from the mean log-likelihood based on k bootstrap samples of size m [8].

The algorithm can be realized exploiting the inconveniently parallel structure and prevents rescan the entire data set in the iterations.

3) Leveraging methods

İn proposed leveraging methods by Ma & Sun one samples a small proportion of the data with concrete weights (subsample) from the full sample. Then performs intended computations for the full sample using the small subsample as a surrogate. The key to success of the leveraging methods is to set up the weights, the non-uniform sampling probabilities [8, 26].

Leveraging methods are different from the traditional subsampling or m-out-of-n bootstrap in that [26]:

1) they are used to achieve realizable computation even if the simple analytic results are available;

2) they enable visualization of the data when visualization of the full sample is unfeasible;

3) they usually use unequal sampling probabilities for subsampling data.

This approach is very unique design in allowing pervasive access to extract information from large volumes of data without the need for high-performance computing.

b. Divide and conquer

This algorithm consists of three stages:

1) divisions a big dataset into K blocks;

2) processes each block separately (possibly in parallel);

3) aggregates the solutions from each block to form a final solution to the full data [8].

For general nonlinear assessment equations were offered a linear approximation of the estimating equations with the Taylor expansion at the solution in each block.

Lin & Xi showed that the aggregated estimator has the same limit as the estimator from the full data [9].

In step 2, penalized regression is applied to each block separately with a sparsity-inducing penalty function satisfying certain regularity conditions. This approach can lead to differential variable selection among the blocks, as different blocks of data may result in penalized estimates with different non-zero regression coefficients.

In the 3rd stage the results of K blocks are combined for creating common evaluation. In this stage the common result of assessments due from combining of confidence distributions' idea in meta-analysis.

c. Sequential Updating for Stream Data

In some situations, the data come in streams or large chunks, and a sequentially updated analysis is desirable without storage demands. Schifano et al. [27] expand upon the work of Lin & Xi [9] in several significant ways.

Firstly, in this method "divide-and-conquer" type variance estimates of regression parameters are introduced in the linear model and estimating equation settings [8, 27].

This variance assessment allow users to make inferences about the true regression parameters based upon previously developed divide-and-conquer point estimates of the regression parameters.

Secondly, they expand iterative assessment algorithms and statistical inferences for linear models and estimating equations (for update as new data arrive) [8, 27].

Thirdly, during dealing with blocks of data are used the issue of possible rank deficiencies and the uniqueness properties of the combined and cumulative estimators during using a generalized inverse by the authors.

In addition, a new online-updated estimator of the regression coefficients corresponding estimator of the standard error in the estimating equation setting which takes advantage of information from the previous data are introduced by authors.

The Expectation-maximization (EM) algorithm has been widely used in scientific computing for parameter estimation in presence of missing data. The successes of the iterative algorithms in modern scientific computing create interest to develop some iterative algorithms that are feasible for big data [6].

Iterative Monte Carlo methods, such as MCMC, stochastic approximation and EM, have proven to be very powerful tools for statistical data analysis. However, their computer-intensive nature, which typically requires a large number of iterations and a complete scan of the full dataset for each iteration, precludes their use for big data analysis. In [13] review of the latest developments of iterative Monte Carlo methods is provided for big data analysis.

The researches show that depending on application fields, different statistical methods are used in the processing of big data. But, standard statistical techniques are usually not well conformed to rule Big Data, and many researchers have proposed extensions of classical techniques or entirely novel methods. For example, C. Angelini and his colleagues have been expanded recently developing functional Bayesian Methods specifically designed for time-course microarray data. The methods successfully deal with various technical difficulties that arise in this type of experiments such as a large number of genes, a small number of observations, non-uniform sampling intervals, missing or multiple data and temporal dependence between observations for each gene. Berk and et al. propose a functional mixed-effects model for estimating the temporal pattern of each gene, which is assumed to be a smooth function [28].

Statistics is the science to gather, arrange, and explain data. With the explosion of "Big Data" problems, statistical computing and statistical learning has become a very hot area in numerous scientific fields as well as marketing, finance, and other business subjects. Computational statistics, or statistical computing, is the interface between statistics and computer science [29].

Author gives a forecast about the development tendency of statistical computing in [30]. Its advance is that technology will impact statistical computing more than other factors. This prediction is based on modern observations of the field over the last 40 years. The technology driving this forecast includes not only hardware, but also the software that ensures the infrastructure for individual and community contact with computers. Whether computer scientists eventually take over this field will depend on how actively statisticians participate. Statisticians interested in statistical computing and its future embodiments will have to engage in cooperative research with computer scientists to continue to have an influence.

Statistical learning refers to a set of tools for modeling and understanding compound datasets. The field encompasses many methods such as the noose and tenuous regression, classification and regression trees, and increasing and promote vector machines. Statistical learning includes building a statistical model for forecasting, or estimating, an output based on one or more inputs. Challenges of this nature appear in fields as miscellaneous as business, medicine, astrophysics, and public policy [31].

This book that is presented by Trevor Hastie and et al. depicts the significant thoughts in the field of statistics, data mining, machine learning, and bioinformatics in a general conceptual framework. This main new edition features many topics, including graphical models, random forests, ensemble methods, least angle regression and path algorithms for the lasso, non-negative matrix factorization, and spectral clustering. İn this book developed generalized additive models, much of the statistical modeling software and environment in R/S-PLUS and etc. [14].

Statistical investigation is widely used for myriad scientific applications in order to analysis and infer from data. An important challenge of any statistical analysis intended at large-scale data is to solve the problems of parallel scalability. Using a series of formulas that permit for single-pass, yet numerically robust, pairwise parallel and incremental updates of both arbitrary-order centered statistical moments and co-moments Bennett J. and his colleagues have built an open source parallel statistics framework that performs principal component analysis (PCA) in addition to computing descriptive, correlative, and multi-correlative statistics [15].

Monotonic regression (MR) is an effective tool for appraising functions that are monotonic with respect to input variables. The MR problem has many applications in oper-

ations research, statistics, biology, signal processing, and other areas. Unfortunately, these accurate algorithms can just solve problems that embrace a relatively small number of observations, and so they cannot insure exhaustive results throughout a possible amount of time when addressing medium- or large-scale problems. A fast and highly accurate approximate algorithm which is offered by Burdakov and et al. called the GPAV was recently developed for efficient solving large-scale multivariate MR problems. An approach, that stretches the application area of the GPAV to surround much larger MR problems, is presented. It is based on segmentation of a large-scale MR problem into a set of moderate-scale MR problems, each solved by the GPAV [16].

Authors in [17] submitted a collection of parallel implementations of statistics algorithm developed as part of a common framework over the previous years. Moment-based statistics (which include descriptive, correlative, and multi correlative statistics, principal component analysis (PCA), and k-means statistics) scale approximately linearly with the data set size and number of processes.

Nearly all the methods that model the statistical characteristics of wavelet coefficients are only suitable for a small data set. Most of them are associated to de-noising, texture analysis, and segmentation, classification, and retrieval algorithms. For big data sets, such as space-temporal remote sensing data sets, also need to model their wavelet coefficients to detect the changing trends, find out the intrinsic mechanisms, and represent the rules of their evolution process. L. Wang et al. in [18] proposed to use the GMM to sign the statistical properties of wavelet coefficients of a remote sensing big data set. This contribution is to calculate the model parameters of a big data set using different aspects or dimensions such as time, spectral bands, scales, and textures.

Ling Song offered a novel multi-resolution cluster detection (MCD) method to equate irregularly shaped clusters in space and derived the multi-scale test statistic on a single cell based on likelihood ratio statistic for Bernoulli sequence, Poisson sequence and Normal sequence. The MCD method compared with single scale testing methods controlling for false discovery rate and the spatial scan statistics using simulation and f-MRI data, more effective for discovering irregularly shaped clusters. The implementation of his proposed method does not demand difficult computation, making it convenient for cluster detection for large spatial data [10].

Hongtu Zhu and et al. offered several classes of spatial regression models including spatially varying coefficient models, spatial predictive Gaussian process models, tensor regression models, and Cox functional linear regression models for the cooperative analysis of huge neuroimaging data and clinical and behavioral data. Their statistical models clearly account for a few stylized peculiarities of neuroimaging data: the availability of multiple piecewise slick regions with unknown edges and jumps and essential spatial correlations. They improved some rapid evaluation procedures to simultaneously estimate the assorted coefficient functions and the spatial correlations. They systematically explored the asymptotic properties (e.g., consistency and asymptotic normality) of the multiscale adaptive parameter estimates too. Their Monte Carlo simulation and real data analysis affirmed the perfect performance of their models in distinctive applications [10].

Big data brings problems to even elementary statistical analysis wherefore the barriers in computer memory and computing time. The computer memory barrier is habitually manipulated by a database connection that extracts data in chunks for processing. It is handled by parallel computing, often expedited by graphical processing units. The open source R packages that break the computer memory limit such as bigglm and bigmemory that help parallel computing [32].

Because of their size and complexity, massive data sets bring many computational challenges for statistical analysis. Overcoming the memory limitation and improving computational efficiency of traditional statistical methods are including to these problems.

Statistical aggregation partitions the entire data set into smaller subsets, compresses each subset into certain low-dimensional summary statistics and aggregates the summary statistics to approximate the desired computation based on the entire data. Results are required to be asymptotically equivalent which are gotten from statistical aggregation. Statistical aggregation is especially helpful to support sophisticated statistical analyses for online analytical processing in data cubes.

7. Conclusions

We are currently living science, engineering and technology widespread area, which provided production of big data flow measurable exabyte and zetabyte. Big Data promise modern levels of scientific revelation and economic value.

Big data has become one of the current and future research directions.

It should be noted that, Big Data will revolutionize many fields, including business, healthcare, the scientific research, public administration, and so on.

The successful application of this technology is also available. A large number of microarray data repositories have been created for gene expression investigation, sophisticated cameras are becoming ubiquitous, generating a huge amount of visual data for surveillance, the Square Kilometre Array Telescope is being built for astrophysics research and is expected to generate several petabytes of astronomical data every year. All the datasets, that called Big data have a large number of dimensions (attributes) and pose significant research challenges for statistical analysis and data mining.

Big data present challenges much further beyond the opportunities of classic statistics, requiring joint workforce with domain knowledge, computing technology, and statistical methods. The importance of the following points in order to properly understand the problem:

1. The complexity of data (collected from different sources and different formats).

2. Noisy data challenge: Big Data generally include different kinds of measurement errors, outliers and missing values.

3. Dependent data challenge: in varied types of current data, such as financial time series and so.

It needs processing of more effective statistical methods and algorithms in solving problems, that are robust to data complexity, noises and data dependence.

As shown in studies expanded options enough traditional statistical methods in this field (for example, bootstraps, split-and-conquer algorithms, subsampling-based methods etc.) were developed and new algorithms (multi-resolution cluster detection (MCD) method, a functional mixed-effects model and etc.) were processed by researchers.

It seems that this field will continue to be the subject of research that researchers often applied.

References

1. Alguliyev R. M. “Big Data” phenomenon: Challenges and Opportunities [Text] / R. M. Alguliyev, M. S. Hajirahimova // Problems of Information Technology. – 2014. – Vol. 2. – P. 3–16.
2. Philip, C. L. Data-intensive applications, challenges, techniques and technologies: A survey on Big Data [Text] / C. L. Philip, C.-Y. Zhang // Information Sciences. – 2014. – Vol. 275. – P. 314–347. doi: 10.1016/j.ins.2014.01.015
3. Soares, S. Big Data Governance. An Emerging Imperative. 1st edition [Text] / S. Soares. – MC Press Online, LLC, 2012. – 368 p.
4. Kambatla, K. Trends big data analytics [Text] / K. Kambatla, G. Kollias, V. Kumar, A. Grama // Parallel and Distributed Computing. – 2014. – Vol. 74, Issue 7. – P. 2561–2573. doi: 10.1016/j.jpdc.2014.01.003
5. Fan, J. Challenges of Big Data analysis [Text] / J. Fa, F. Han, H. Liu // National Science Review. – 2014. – Vol. 1, Issue 2. – P. 293–314. doi: 10.1093/nsr/nwt032
6. Jordan, J. M. Statistics for big data: Are statisticians ready for big data? [Text] / J. M. Jordan, D. J. Lin // International Chinese Statistical Association Bulletin. – 2014. – Vol. 26. – P. 59–66.
7. Yu, B. Let us own data science [Text] / B. Yu // IMS Bulletin Online. – 2014. – Vol. 43, Issue 7.
8. Chun W. C. A Survey of Statistical Methods and Computing for Big Data [Electronic resource] / W. C. Chun, M. H. Chen, E. Schifano, J. Wu, J. Yan. – 2015. – Available at: http://de.arxiv.org/abs/1502.07989v1
9. Lin, N. Aggregated estimating equation estimation [Text] / N. Lin, R. Xi // Statistics and Its Interface. – 2011. – Vol. 4, Issue 1. – P. 73–83. doi: 10.4310/sii.2011.v4.n1.a8
10. Chen, M. H. Statistical and Computational Theory and Methodology for Big Data Analysis [Electronic resource] / M. H. Chen, R. Craiu, F. Liang, C. Liu. – Available at: https://www.birs.ca/workshops/2014/14w5086
11. Liang, F. A resampling-based stochastic approximation method for analysis of large geostatistical data [Text] / F. Liang, Y. Cheng, Q. Song, J. Park, P. Yang // Journal of the American Statistical Association. – 2013. – Vol. 108, Issue 501. – P. 325–339. doi: 10.1080/01621459.2012.746061
12. Liang, F. A bootstrap Metropolis–Hastings algorithm for Bayesian analysis of big data. Tech. rep. [Text] / F. Liang, J. Kim. – Texas A & M University, 2013.
13. Liang F. Advanced Markov chain Monte Carlo methods: learning from past samples [Text] / F. Liang, C. Liu, R. J. Carroll. – Wiley, New York, 2010. – 378 p.
14. Hastie, T. The Elements of Statistical, Learning: Data Mining Inference and Prediction. Second edition [Text] / T. Hastie, R. Tibshirani, J. Friedman. – Springer, 2009. doi: 10.1007/978-0-387-84858-7
15. Bennett, J. Numerically stable, single-pass, parallel statistics algorithms [Text] / J. Bennett, R. Grout, P. Pebay, D. Roe, D. Thompson // Proceedings of the IEEE International Conference on Cluster Computing and Workshops, 2009. – P. 1–8. doi: 10.1109/clustr.2009.5289161
16. Sysoev, O. A segmentation-based algorithm for large-scale partially ordered monotonic regression [Text] / O. Sysoev, O. Burdakov, A. Grimvalla // Computational Statistics and Data Analysis. – 2011. – Vol. 55, Issue 8. – P. 2463–2476. doi: 10.1016/j.csda.2011.03.001
17. Pébay, P. Design and performance of a scalable, parallel statistics toolkit [Text] / P. Pébay, D. Thompson, J. Bennett, A. Mascarenhas // Proceedings of the International Symposium on Parallel and Distributed Processing Workshops and Phd Forum, 2011. – P. 1475–1484. doi: 10.1109/ipdps.2011.293
18. Lizhe, W. Estimating the Statistical Characteristics of Remote Sensing Big Data in the Wavelet Transform Domain [Text] / L. Wang, Z. H. Hui, R. Ranjan, A. Zomaya, P. Liu // IEEE Transactions on Emerging Topics in Computing. – 2014. – Vol. 2, Issue 3. – P. 324–339. doi: 10.1109/tetc.2014.2356499
19. Jin, X. Significance and Challenges of Big Data Research [Text] / X. Jin, B. W. Wah, X. Cheng, Y. Wang // Big Data Research. – 2015. – Vol. 2, Issue 2. – P. 59–64. doi: 10.1016/j.bdr.2015.01.006
20. ITU-T Technology Watch [Electronic resource]. – Big data: Big today, normal tomorrow, 2013. – Available at: http://unstats.un.org/unsd/trade/events/2014/beijing/documents/other/ITU.pdf
21. Big Data, Big Impact: New Possibilities for International Development [Electronic resource]. – Available at: http://www.weforum.org/reports/big-data-big-impact-new-possibilities-international-development
22. 8 top challenges big data brings to statisticians [Electronic resource]. – Available at: http://www.fiercebigdata.com/story/8-top-challenges-big-data-brings-statisticians/2014-07-14
23. 2013 International Year of Statistics [Electronic resource]. – Available at: http://www.isi-web.org/recent-pages/490-2013-international-year-of-statistics
24. Kleiner, A. A scalable bootstrap for massive data [Text] / A. Kleiner, A. Talwalkar, P. Sarkar, M. I. Jordan // Journal of the Royal Statistical Society: Series B (Statistical Methodology). – 2014. – Vol. 76, Issue 4. – P. 795–816. doi: 10.1111/rssb.12050
25. Basiri, S. Robust, scalable and fast bootstrap method for analyzing large scale data [Electronic resource] / S. Basiri, E. Ollila. – 2015. – Available at: http://arxiv.org/pdf/1504.02382
26. Ma, P. Leveraging for big data regression [Text] / P. Ma, X. Sun // WIREs Computational Statistics. – 2015. – Vol. 7. – P. 70–76. doi: 10.1002/wics.1324

27. Schifano, E. D. Online updating of statistical inference in the big data setting. Tech. Rep. [Text] / E. D. Schifano, J. Wu, C. Wang, J. Yan, M.-H. Chen. – University of Connecticut, Storrs, Connecticut, 2014.
28. Advanced Statistical Methods for the Analysis of Large Data-Sets [Text] / A. Di Ciaccio, M. Coli, J. M. Angulo Ibanez (Eds.). – Springer, 2012. doi: 10.1007/978-3-642-21037-2
29. Computational statistics [Electronic resource]. – Available at: https://en.wikipedia.org/wiki/Computational_statistics
30. Wilkinson, L. The future of statistical computing [Text] / L. Wilkinson // Technometrics. – 2008. – Vol. 50, Issue 4. – P. 418–435. doi: 10.1198/004017008000000460
31. James, G. An Introduction to Statistical Learning with Applications in R [Electronic resource] / G. James, D. Witten, T. Hastie, R. Tibshirani. – AVailable at: http://www-bcf.usc.edu/
32. Schbidberger, M. State of the art in parallel computing with R [Text] / M. Schbidberger, M. Morgan, D. Eddelbuettel et. al. // Journal of Statistical Software. – 2009. – Vol. 31, Issue 1. – P. 1–27.

Розглянуто розпізнавання стану структури прихованої частини складних мережевих об'єктів в умовах обмеженої інформації від їх важкодоступних елементів. Метод розпізнавання стану мережних об'єктів ліг в основу побудови інтелектуальної системи підтримки прийняття рішень при експлуатації та реінжинірингу відновлюваних бездротових комп'ютерних мереж з недоступними для моніторингу елементами

Ключові слова: штучний інтелект, зорове відображення, мережеві структури, важкодоступні елементи

Рассмотрено распознавание состояния структуры скрытой части сложных сетевых объектов в условиях ограниченной информации от их труднодоступных элементов. Метод распознавания состояния сетевых объектов лег в основу построения интеллектуальной системы поддержки принятия решений при эксплуатации и реинжиниринге восстанавливаемых беспроводных компьютерных сетей с недоступными для непосредственного мониторинга элементами

Ключевые слова: искусственный интеллект, зрительное отображение, сетевые структуры, труднодоступные элементы

УДК 004.08:005.8
DOI: 10.15587/1729-4061.2015.51186

ПЕРЕТВОРЕННЯ СТРУКТУРИ СКЛАДНОЇ ТЕХНІЧНОЇ СИСТЕМИ ІЗ ЧАСТКОВО НЕДОСТУПНИМИ ЕЛЕМЕНТАМИ ДО ЗОРОВОГО ОБРАЗУ

С. А. Нестеренко
Доктор технических наук, профессор*
E-mail: san@opu.ua
А. О. Становський
Кафедра компьютерных интеллектуальных систем и сетей*
E-mail: redline@normaplus.ua
А. В. Торопенко
Кандидат технических наук
Кафедра нефтегазового и химического машиностроения**
E-mail: alla.androsyk@gmail.com
П. С. Швець
Кандидат технических наук
Кафедра электроснабжения и
энергетического менеджмента**
E-mail: pshvets@mail.ru
*Кафедра компьютерных
интеллектуальных систем и сетей**
**Одесский национальный политехнический университет
пр. Шевченко, 1, г. Одесса, Украина, 65044

1. Вступ

Будь-яка система розпізнавання надійності пошкоджуваних під час зберігання та експлуатації мережевих об'єктів, тобто об'єктів, які складаються з окремих елементів та зв'язків між ними, потребує, як мінімум, відомостей про початковий стан їхньої структури, а також результатів аналізу структури поточного стану. Якщо подібні об'єкти спочатку або в результаті пошкоджень частково *недоступні* для моніторингу, з таким аналізом виникають проблеми [1, 2].

В цьому випадку дослідник має змогу отримати лише обмежену інформацію про значення деяких характеристик *доступної* частини мережевого об'єкта за деякий період до поточного часу включно. Це можуть бути вимірювані на доступній частині часові тренди параметрів стану термодинамічних систем (температура, тиск, концентрація, тощо), механічних характе-