Correlation analysis and causal analysis in the era of big data

Xu Gejingting¹, Jiang Ruiqiong² Wei Wei², Jiang Libao¹ and Yue Zhenjun¹

¹the College of Communications Engineering, the Army Engineering University of PLA, China
²31106 Troops, PLA, China

Abstract. Based on the essence of correlation analysis, this paper studies the difference and relation between correlation analysis and causal analysis. It points out that in the era of big data, correlation analysis and causal analysis are different aspects of the analysis of things. They are neither antagonistic nor mutually replaceable but a dependency. Big data provides better and more powerful functions for correlation analysis. Correlation analysis provides direction and verification path for causal analysis. Causal analysis provides deep support for correlation analysis. Studying the application methods of correlation analysis and causal analysis under the condition of big data.

1. Introduction

Professor Viktor Mayer-Schönberger¹, known as the first data scientist to have insight into the development trend of big data and one of the most respected authoritative spokespeople of Oxford University, points out in the book "Big Data"² that with the advent of the big data, the traditional way of thinking will undergo tremendous changes and data-based correlation analysis will replace causal analysis and become the main method of object analysis. This view has been supported by many scholars, but there is also controversy.³ Zhou Tao, the Chinese translator of the book "Big Data", believes that "Abandoning the pursuit of causality means giving up the intellectual superiority of human beings above computers. It is the indulgence and degeneration of human beings." This paper argues that correlation analysis and causality analysis are not antagonistic. On the one hand, “correlation” replaces “causality” is not a new topic. In traditional study of statistical data analysis and data mining, such cases are everywhere. On the other hand, replacing causal analysis with correlation analysis in certain special fields and specific applications can indeed achieve greater benefits. In the third aspect, the improvement of the status of correlation analysis caused by big data is not the exclusion and boycott of causal analysis, but the strengthening and support of causal analysis methods, which can open up new ways for causal analysis. This paper summarizes the role of big data in analysis by discussing about correlation analysis and causal analysis and the relationship between them under the condition of big data.

2. Causal analysis and correlation analysis

2.1. Significance of analysis

The relationship between things is the objective existence. Therefore, the analysis of the relationship between things is an important part of the research work. The analysis of the relationship between things helps to strengthen people's depth of understanding, reduce the uncertainty of the development of things and improve the quality of decision-making. Secondly, it can help people to judge the related things and
improve the level of thinking. The third is to help people speculate something that is difficult to see according to the law of evolution of something which is easy to find. Causality and correlation are the two most commonly used scales when people do relationship analysis. Historically, due to the superstition of classical mathematics and the worship of logical thinking, people pay more attention to causal analysis than to correlation analysis. In recent years, the advent of big data and big data technology has brought new vitality and successful application paradigms to correlation analysis, leading people to the other extreme——thinking that correlation analysis is enough to replace causal analysis. In fact, they are two angles and two methods and they complement each other and cannot be neglected.

2.2. Overview of causality
The causality reveals the relationship between two things through the material mechanism and reflects the inherent physical properties between things. There is a physical dependence between objects with causality. Causal analysis is an attempt to engage in the intrinsic properties of things to discover the physical connection of things. Mastering the causal relationship between things is equivalent to mastering the inevitable law of development and change between things and thus making correct decisions. The causal relationship is deterministic and the mathematical model that characterizes causality is a deterministic mathematical function.

2.3. Overview of correlation
Correlation can be divided into two types: generalized correlation and narrow correlation. As long as two things change at the same time, it can be regarded as generalized correlation. As correlation need vary in degree, in order to measure correlation from a quantitative perspective, narrow correlation is introduced: When things can be represented by real numbers, linear correlation is used as a criterion for correlation and the correlation coefficient indicates the degree of correlation and there are other ways such as rank correlation and so on.

Different from causality, correlation is the external performance level of the object and reflects the similarity between the phenomena displayed in the process of the development. Causal analysis can help us to identify key clues in numerous phenomena and reveal objective facts which are not easily found. Especially in business, optimizing the placement of merchandise in supermarket and directional commercial advertising based on the relevance of customer shopping can help to improve sales efficiency. This is also proved by facts.

Another point that differs from causality is that the correlation has probability meanings, that is to say, the law revealed by the correlation is only established for partial data or probability. If the correlation coefficient is large, it may establish with high probability (but the correlation coefficient is not the probability), the application of the conclusion will be wider. While even two things with a correlation coefficient of 1, there have exceptions. Therefore, decisions based solely on the conclusions of the correlation analysis are only correct in the sense of probability. It means that they can only be demonstrated if they can be applied in large numbers of occasions (such as retail).

2.4. The relationship between causality and correlation
We know that the external manifestation of things is governed by the internal mechanism and the environmental conditions. Therefore, there are many possibilities for reflecting the internal characteristics of things corresponding to the correlation of external performance. Specifically, things with causality necessarily have correlation and things with partial causality also show correlation. A typical example is that the correlation between the height of children and their parents. Two different results of the same thing also have correlation. For example, there is no causality between soybean yield and corn yield (they are the same season crops, all related to meteorological conditions), but they show strong correlation. Since modern computer-based correlation analysis is more often by means of data analysis, some completely unrelated data will also show some correlation. Therefore, the relationship between two things with correlation still requires causal analysis to confirm or negate some conclusions.
It is considered that causal analysis and correlation analysis are two complementary angles to analyze and they are not antagonistic.

3. Correlation Analysis in the era of big data

3.1. Concepts and methods of Correlation Analysis

Correlation analysis is a statistical method which infers the correlation between variables by observing the interdependency of the trend of data. The textbooks of public course in science and engineering university usually introduce linear correlation analysis. There are other types of correlation analysis in more professional textbooks or literature. This paper discusses linear correlation analysis as an example.

Assume \( X = \{x_1, x_2, ..., x_n\} \), \( Y = \{y_1, y_2, ..., y_n\} \) are the recorded value of two factors (such as height and weight), in order to study its correlation, define the correlation coefficient:

\[
\begin{align*}
\rho_{xy} &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2} \sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \\
\bar{x} &= \frac{1}{n} \sum_{i=1}^{n} x_i, \quad \bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i
\end{align*}
\]

The meaning of the formula is that: \( X \) and \( Y \) are regarded as two points in the \( n \)-dimensional space (representing the state of the two indicators in the same space). We can obtain two rays by respectively connecting them to the origin. \( r_{xy} \) is the cosine of the angle of the rays. The closer the cosine value is to 1, the higher the degree of coincidence of the two rays are, which also means the more regular the change of the two quantities are. In extreme cases, when \( r_{xy} = 1 \) or \(-1\), there is a constant \( k, b \), so that

\[ Y = kX + b \]

This means that there is a linear relationship between \( x \) and \( y \). To further illustrate the characteristics of the correlation coefficients, we have selected the data of student transcript, which is showed in Table 1.

| English | Chinese | Mathematics | Computer | Physical education | Society |
|---------|---------|-------------|----------|--------------------|---------|
| 1       | 80      | 96          | 96       | 96                 | 34      | 96      |
| 2       | 70      | 84          | 84       | 84                 | 12      | 84      |
| 3       | 60      | 72          | 72       | 72                 | 78      | 72      |
| 4       | 50      | 60          | 60       | 60                 | 56      | 56      |
| 5       | 40      | 48          | 24       | 36                 | 22      | 22      |

Analyzing the correlation between English, Chinese, mathematics, computer, physical education and society, we show the correlation coefficient in Table 2.

Because of a strict linear relationship between English and Chinese scores (The score of Chinese test is 1.2 times higher than English test), so the correlation coefficient between the two is 1. The more likely the data is to linear relationship, the closer the correlation coefficient is to 1.

Therefore, the characteristics of the correlation analysis can be summarized as below.

(1) The correlation only reflects the linear interdependency of the law of numerical variation between the two objects. For the nonlinear interdependency variables, it can not reach the true conclusion.

(2) Whether there is correlation between variables and the degree of the correlation depend only on the actual observations of the variables and have nothing to do with the physical meaning.

(3) Because of (2), different observations will lead to different conclusions of the correlation.

(4) As a result of (3), the conclusions of the correlation will only be valuable in the sense of probability.

(5) Because of (4), decisions based on correlation analysis are risky.
Table 2. Correlation coefficient.

|     | English | Chinese | Mathematics | Computer | Physical education | Society |
|-----|---------|---------|-------------|----------|--------------------|---------|
| 1   | 1       | 1       | 1           | 1        | 1                  |         |
| 2   | 0.961524 | 0.961524 | 1           | 1        | 1                  |         |
| 3   | 0.986394 | 0.986394 | 0.993605    | 1        | 1                  |         |
| 4   | -0.11861 | -0.11861 | 0.068429    | -0.0078  | 1                  |         |
| 5   | 0.969436 | 0.969436 | 0.998717    | 0.996089 | 0.057493           | 1       |

Due to the various defects above, people have imposed various restrictions on the practical application of the correlation analysis, such as making a significant test on the correlation coefficient.

3.2. The value of Correlation Analysis in the ear of big data

The advent and development of big data and big data technology has injected new vitality into correlation analysis and makes it more usable.

First of all, from the perspective of thinking mode, data has become the greatest wealth and inevitably affects people's way of life and thinking in big data era, while "data thinking" gradually achieve the same status as logical thinking. The more cases you see, the more people trust their eyes, so the logical judgment of the brain becomes more and more lazy. This is called data thinking. It is the absoluteization of traditional inductive thinking and the simplification of the thinking process from partial facts to general conclusions. The traditional inductive thinking is the fact that the data is reflected. People summarize the general laws of the object through the role of the brain. Therefore, the more perfect the data get, the closer it will be to the law. With sufficient data or even the full data, people have reasons to believe that these data contain the evolution of the objects and correct decisions can be made. However, in the data thinking, the induction process is handed over by computer. While the efficiency and speed are improved, the logical thinking process disappear.

Secondly, the development of computing technology provides technical feasibility for the massive calculations for data analysis. The traditional statistical analysis belongs to sampling statistics, that is, using partial samples to infer the overall characteristics, which is inevitably limited to the predicament of partiality. Although theoretically, people impose various restrictions on the sampling process to make the sampling data fully representative. However, in practice, it is still impossible to make the sampled data reflect all the characteristics and other assumptions must be added, such as the overall obeying the normal distribution and so on. The assumptions are somewhat small-minded, because the correlation analysis of big data uses all the data, which includes all the information. No additional assumptions need to be made.

Thirdly, causal analysis relies on human logical thinking ability while correlation analysis relies on the computing power of computers. Compared with the human brain, computer reasoning method is simple, fast and easy to implement. Therefore, it is used in retail business, where automatically calculated is needed and errors do not lead to fatal losses. So application correlation analysis is naturally more easily recognized than causal analysis.

Finally, in the business world, people's decisions are directed at high-probability events. In other words, if the records show that 30% of customers buy coffee while buying bread and only 20% buy tea while buying bread at the same time, it is enough to show that while buying bread, coffee have a stronger correlation than tea. Designing a shelf configuration or doing market strategy according to this finding will greatly increase sales without considering the reasons behind. The superficial effect makes significant economic benefit and also masks the utility of causal analysis to a certain extent. Therefore, it is not strange that the view that correlation analysis can replace the causal analysis appears.

4. The Combination of Correlation Analysis and Causal Analysis under the condition of big data

4.1. The characteristics of correlation Analysis and causal Analysis
Causal analysis is to analyze the relationship between things from the mechanism level and to discover the inevitability of the development. Its advantage lies in the ability to grasp the inherent laws and accurately predict the future trend. Correlation analysis speculate the law of the development through external manifestation. The accuracy and credibility of the inference depends naturally on the amount of data and the analysis methods. Sufficient data and rich analysis methods make up its shortcomings based on big data. Correlation analysis is simple, fast and easy to implement, so it is more suitable than other methods in the case of real-time decision. It has been more widely used in business, society and other occasions than causal analysis, so that experts and scholars advocate the use of correlation analysis methods to replace causal analysis. We believe that although the purpose of correlation analysis and causal analysis is to reveal the relevance of the development of things, to try to monitor another thing through the analysis of one object, but the different foothold of them determine their different application occasion. The example in business about retail we write above is based on correlation analysis, which pays attention to “what” and do not care about “why”. However, we should be noticed that, it can only be understood as “coincidence” without causality. The decisions made based on this “coincidence” are only useful when the risk is small. On a life-and-death occasion such as competitive intelligence analysis, military intelligence analysis, etc., the basis for decision must consider causal analysis.

4.2. The advantages of correlation analysis under the condition of big data

However, correlation analysis under the condition of big data provides a technical guide for causal analysis with its convenience and speed. In a complex world, it is sometimes difficult to directly perform causal analysis. Do correlation analysis first by computer can easily indicate the direction of causal analysis. As mentioned earlier, the correlation shown by the data may come from direct causality, indirect causality, partial causality and completely random “coincidence”. But no matter which relationship it is, further causal analysis on this basis is much more efficient than causal analysis aimlessly.

On the other hand, correlation analysis can also be used as a test for causal analysis. When causal analysis shows that there is a causal relationship between the two factors and the data indicates that it is irrelevant, we need to examine carefully whether the analysis process is wrong or whether other factors that have not been considered.

4.3. Conclusion

In a word, the emergence of big data and big data technology has greatly expanded the functions of correlation analysis, enhanced its status and has successfully applied in business, sociology and other fields. However, due to its inherent characteristics of randomness and superficiality, it can not be used independently in high risk areas such as competitive intelligence analysis and military intelligence analysis. However, using the easy operation of correlation analysis in big data era to discover or verify certain facts or phenomena that are not easily found can also bring huge benefits. In intelligence analysis, doing correlation analysis before causal analysis can reduce the complexity of causal analysis and improve the efficiency. After the causal analysis came to a conclusion, correlation analysis is equivalent a verification method for causal analysis which can improve the credibility.

References

[1] http://baike.baidu.com/link?url=RVE2WTWU9H AeXzq3Q30CwttwTHThYJKDeRWsG0A2lixZ RjEBIxQsgebMD6W7CQ_HUl YcvAVk_cB6eVGgW0fFmEq.
[2] Victor Meyer Schonberg: "Big Data", Zhejiang People's Publishing House, 2012.12.
[3] Gu Hongwen: Causality and Relevance in “Big Data”,http://www.china-cloud.com/yunhudong /yunzhuannan/guhongwen/2013/1201/21934.html
[4] Liang Jiyue, Feng Chenjiao, Song Peng: A Survey on Correlation Analysis of Big Data , Chinese Journal of Computers, 2016(1):1-18.