Research on the Application Development of Data Science in Customer Segmentation

Yuxuan Xu
Department of Economics, New York University, New York, United States of America
email:yx1151@nyu.edu

Keywords: Data science, Customer segmentation, Application development

Abstract. Data science is an emerging discipline that extracts knowledge from data, and is formally put forward in the context of the "fourth paradigm" of scientific research. The data used for business decision-making and scientific research have radically changed, characterized by the original mainstream sampling, structured, small-scale development to full data, semi-structured and unstructured, large-scale, driven data science developed into both Including traditional statistical methods, but also includes data mining, text mining, process mining and large data and other emerging technologies interdisciplinary. Customer segmentation is a typical data oriented business and research areas, data science in one of the application shows that the subject contains a variety of methods, can successfully extract complex customer related data contains information and knowledge, can solve the problem of the feasibility and effectiveness of customer segmentation, accuracy, and business practices for customer segmentation study provides good technology support and decision support.

Introduction

Data generation, collection, storage and handling of evolution has given rise to data explosion era, which needs a specializing in various types of data, status, properties, form of organization, changing mode and law science, reveals the nature and human behavior phenomenon and law, provides a new method for scientific research and business intelligence. As a result, the concept of data science has been proposed, and it is rapidly becoming more and more widely used in business and research.

Customer segmentation refers to the customer according to customer's attribute set. As one of the most important strategic resource of enterprise, customer relationship between the satisfaction of their needs the enterprise survival, growth and sustainable development, in order to satisfy the heterogeneity of customer needs, formulate corresponding differential management strategy, enterprise widely used customer segmentation theory as an important management tool, the tool even has become a global one of the top ten management tool in the highest utilization rate. Effective customer data analysis is the key to customer segmentation success. In the practice of the customer segmentation, factors that limit the role of the is no longer the customer data information is insufficient, but the potential value of the customer information resources mining is inadequate. Data science can handle various types and the size of the data, make full use of the data contains information extracted is conducive to business decisions and scientific research knowledge, very suitable for customer segmentation this depend on the data of field.

Data science and the fourth paradigm

Data Science was originally called datalogy. It was first proposed by Peter Naur in 1966. In the International Federation of Classification Societies (IFCS) International conference, the term Data Science first appeared in the title of the conference (Data Science, Classification, and related methods). The concept of data science was widely regarded in academic circles in the 20th century. In 2002, the international scientific council, the data commission, began publishing data science journals. In 2003, the university of Columbia began publishing data science magazine, which covered the application of statistical methods and quantitative research.

Copyright © 2017, the Authors. Published by Atlantis Press.
This is an open access article under the CC BY-NC license (http://creativecommons.org/licenses/by-nc/4.0/).

As you can see from the above points, data science is a new interdisciplinary discipline, and its real development is at a stage of rapid development. In fact, the fourth paradigm, the idea of data intensive science, has greatly contributed to the formal development of data science. In the mid 1990s, the late Turing award winner, Gray (Jim Gray) [1] proposed the scientific research of "the fourth paradigm" - data-intensive science. Differing from that of the experiment, theory, calculation of the three paradigms, "the fourth paradigm", the need to "will be used for data calculation, rather than the data used to calculate", this view is, in fact, the data alone distinguishes science from computing science.

Data Evolution Drives Data Science Generation and Development

As mentioned earlier, the data processing and analysis method of data science are developed respectively in different disciplines, including mathematics, statistics and information technology in the field of multiple technologies and theories, such as: signal process, stochastic models, machine learning, statistical learning, computer programming, data engineering, pattern recognition and learning, visualization, forecasting techniques, uncertainty model, the data warehouse and high-performance computing. Data science is particularly concerned with scaling up to big data, but it is widely believed that data science is limited to big data. In a word, in order to extract more information and knowledge from data, many new technology with the evolution of the data characteristics, have been incorporated into the category of scientific data.

Statistics. Just as some researchers think that data science is developed on the basis of statistics. Statistics are the first class of technologies involved in data science. Before 1900, statistics are not as an independent discipline, and the processing of data is mainly a nationwide census register, and generally is a few simple data summary and comparison. During the period between 1920 and 1960, the focus of statistics was gradually condensed into small data, which produced the classic statistical method of statistical inference. Subsequently, statistical methods were widely used and developed rapidly. Statistics processing data are sampled, structured, and relatively small.

The Development of Data Science Outside of Statistics. The real sense of modern statistics is from the processing of small data, imperfect experiments such as the development of such practical problems, and data science is due to another kind of real problems and the rise - all data, semi-structured and non-structural, The information contained in large-scale records needs to be fully exploited to produce greater value.

Full Data: Data Mining. With the arrival of the data age, for the need to study the problem can often get the overall data, data collection changes directly down the significance of the sample. The core of the modern statistical analysis method is the sampling inference, that is, how to infer the population in the case of observing the sample. In the case of full data, however, the inferences have lost their value. As a result, it is preferable to use data for all of the data available. Data mining as a kind of method that can handle the whole data, in many cases, makes it easier to model discovery, and thus become a kind of important method in data science. After data mining, another can handle all data of technology is a method that big data is not random analysis such shortcuts, and use all the data for analysis.

Multiple Types: Text Mining and Process Mining. Although the goal of all kinds of data processing technologies is to extract information and knowledge from data, the range of data that these technologies can handle is different. The data itself is a broad concept, including structured, semi-structured, and unstructured. Although statistical research data type is rich, it needs to be structured in the early stage of the process. For example, a statistical category, orderly data used by qualitative data, such as distance data, constant ratio data, such as quantitative data, are structured data. Process of text mining, and other technology can process data in text, operation records and other form of social information, location information, behavior habit, preference information and other dimensions of information. Thus, the maximum use that before can't use the records is to analyze human behavior data, enables the analysis of the data range expanding fast. Data science by incorporating these new data processing technology, can make a wider range of data and includes not only the traditional structured data, but also include statistical unable to process the text, images,

video, audio, web logs, and other non structured and half structured data. Overall, data science is much less demanding in data structures.

Large Scale: Big Data. As technology advances, humans began to take measurements to record all can record the data. Thus, the data size increases sharply, and the data quantity achieved from GB level to PB level. Big data is also known as huge amounts of data (bigdata, megadata), which presents the characteristics of 4V: volume, velocity, variety and value. It need s to deal with the new model that can have better decision-making, insight, and process optimization ability of mass, high rate of growth and diversification of information assets.

Since it is not possible to determine which data is absolutely useless, the risk of selecting partial data for deletion is greater than the cost of storing large amounts of data. Since it is difficult to reduce the amount of data by deleting, the traditional relational database does not have the flexibility and scalability to cope with the rapid expansion of data volume, nor the ability to query, calculate and statistically analyze large data quickly and flexibly. Thus, the new technology is needed to put forward to deal with big data. Big data contains higher efficiency storage technology and has the significant scalability and extensibility, which can well adapt to the rapid expansion of the amount of data, be infinitely expand through simple increase in computer storage capacity. In addition, it includes advanced query language, data format with flexibility and adaptability of reality (without a fixed format), and can be within the scope of tolerable time to complete a series of data processing. Because big data analysis to solve other technology has failed to provide solutions to mass data storage and computational feasibility and effectiveness of the problem, the science and technology is full to the data.

Data Oriented: Algorithms and Models. Data-oriented in the fourth paradigm of the wave, has been agreed by more and more researchers and managers. But the data orientation still has two kinds of sounds based on the model and the algorithm. The model based method assumes that there is some kind of generation mechanism behind the data. The basic view is that the model obtained is not only correct but also accurate. The algorithm based method realizes that the complex real world can not be characterized by the mathematical formula. For the complex, high-dimensional, non-linear data sets in reality, the specific mathematical model is not set, the function mechanism is not discussed, Do the corresponding restrictive assumptions. In many applications, the algorithm model is given a solution to a specific problem, not a statistical solution. The explanability of the algorithm model is weak, but the restraint of the data distribution structure is assumed to be less, and the computational efficiency has a great advantage and the scalability is stronger.

Since the mid 1980s, the algorithm model with the rapid development of computer technology and the rapid growth, however, is largely outside the field of statistics "quietly", such as artificial neural network, support vector machines, decision tree, machine learning and data mining methods such as random forests. The algorithm model is more and more widely appreciated by the academic community in its natural and computer compatibility. Data-driven data analysis is an important trend that cannot be avoided.

Customer Segmentation based on Data Science

Customer segmentation has natural ties with the data. Accurately segmenting customers need to rely on data, at the same time, enterprises under the current technical environment can provide a lot of data about customer. Using science and technology of the data, can make full use of the data from different angles in the information. Three-dimensional fully draw the outline of the various features of each kind of customer group. At present, the typical data science and technology used in customer segmentation are clustering analysis, artificial neural network, text mining, etc. The application of these technologies can achieve different content, dimensions and focus on customer segmentation, and the conclusions of the business practice is very good reference function.

Clustering Analysis. Clustering analysis is a data mining technology that divides data objects into several classes or clusters according to their characteristics. A cluster is a set of data objects. Objects in the same cluster are similar to each other, and objects in different clusters are different

from each other. Many studies have applied clustering analysis to customer segmentation studies. Ferreira Lope [2] emphasizes the advantages of joint analysis and clustering analysis in customer segmentation, through cluster analysis to understand consumer preferences, and accordingly grouped customers to develop more targeted marketing strategies. Simunaniemi AM [3] and so on use semi-structured questionnaire survey method and two-step clustering analysis of consumer eating habits analysis to confirm that the cluster analysis can be targeted Sexual health and nutrition guidance for the consumer group. O. Dzobo [4] introduces the value of customer segmentation in the field of power system, and uses the hierarchical clustering technology to subdivide the customers from the economic scale, economic activity and energy consumption of the power industry. Henriette Müller [5] establishes a multi-dimensional customer segmentation model for the stability analysis of the power system. The three variables of scale, economic activity and energy consumption are used to subdivide the customers according to their load characteristics.

Artificial Neural Network. Artificial Neural Network is a similar to the brain neural network structure and function of the mathematical model. It is a series of processing units using the appropriate way to interconnect a non-linear information processing system. This method is an artificial intelligence algorithm, which is also a kind of data mining technology, and has the characteristics of adaptability, self-organization and fault tolerance. It can realize the intelligent analysis quickly and accurately, and make the forecast and evaluation in the future. Recognition, data processing and automation control and other fields have achieved good results.

Many studies have applied artificial neural network methods to customer segmentation. Ali, J. and Rao, C.P. [6] explore the more effective method of market segmentation on the basis of continuous improvement of information processing and communication technology, and elaborates the feasibility of neural network model in detail. Kauko, T. [7] uses two kinds of neural network model - self-organizing map (SOM) and learning vector quantization (LVQ) model of Finland Helsinki's real estate market segmentation and finds that customers are more concerned about the geographical location and housing types, and house prices are less considered factors. Derrick S. Bullone and Michelle Roehm [8] apply fuzzy artificial neural network analysis technology to test membership clustering criteria, based on the existing method to determine the target market segmentation, and verify the advantages of different market segments. Velu, C.M. and Kashwan, K.R. [9] use the artificial neural network intelligent model to study the customer's consumer behavior and customer classification.

Text Mining. Text mining, also known as text data mining or text knowledge discovery, refers to the process of extracting unknown, understandable, and ultimately available knowledge from a large number of textual data, while using the knowledge to better organize the information. Intuitively, when the object of data mining entirely by the text of this data type, this process is called text mining. The basic idea of text mining is to first cut the text, the information extraction, the use of unstructured text data classification, clustering, association analysis technology into structured data, and then based on structured data to find information between the relationship, and trend forecasting. Text mining as a hot field of data mining has been widely concerned by government, business and research institutions, but the method applied to customer segmentation research has just started with significant development potential.

Many researchers have applied the text mining technology to customer segmentation, focusing on the customer's attitude, emotion, viewpoint and so on. Sun [10] has pointed out that it is becoming increasingly challenging to provide appropriate personalized push content to customers such as YouTube and Flicker and other social media sites, and text mining can effectively deal with this challenge. Oded Netzer [11] proposed that in the era of Web2.0, a large number of text data formed by the blog, forum, chat tool records of user ideas, beliefs and experience, can be analyzed throghy the text mining method. The attitude of the opponent's products will be transformed into market structure and competitive landscape-related information.

Conclusion

Data science is an emerging discipline that extracts knowledge from data, and is formally proposed

in the context of the rise of the "fourth paradigm" of scientific research. For business decisions and scientific research data produced fundamental change, its features from the mainstream of development to full data sampling, structured, small-scale, semi-structured, and unstructured, large-scale, drive data science development become contains both traditional statistical method, and data mining, text mining, mining process and emerging technology such as large data of cross subject. At the same time, the data science also reflects the trend from the weighted model to the weighted algorithm.

Customer segmentation is a typical data oriented business and research areas. Data science in one of the application shows that the various methods included in this discipline can successfully extract the information and knowledge contained in the complex customer-related data, and can solve the feasibility, validity and accuracy of customer segmentation, and provide technical support and decision support for customer segmentation research and business practice.

References

[1] T Hey, S Tansley, K Tolle. The Fourth Paradigm: Data-Intensive Scientific Discovery [C]. Microsoft Research, 2012: 177-185.

[2] Ferreira Lopes, SérgioDominique ,RialBoubeta, Antonio, Varela Mallou, Jesús. Post Hoc Tourist Segmentation with Conjoint and Cluster Analysis[J]. PASOS : Revista de Turismo y Patrimonio Cultural, 2009: 73.

[3] Simunaniemi A-M,NydahlM,Andersson A. Cluster analysis of fruit and vegetable-related perceptions: an alternative approach of consumer segmentation[J]. Journal of Human Nutrition and Dietetics,2013: 261.

[4] Henriette Müller,Ulrich Hamm. Stability of market segmentation with cluster analysis - A methodological approach[J]. Food Quality and Preference,2014: 34.

[5] Dzobo,K. Alvehag,C.T. Gaunt,R. Herman. Multi-dimensional customer segmentation model for power system reliability-worth analysis [J]. International Journal of Electrical Power and Energy Systems,2014,62.

[6] Ali, J; Rao, CP, Neural networks model: A viable approach for micro market segmentation,[J]. Summer Marketing Educators Conference of the American-Marketing-Association, 2010:320-321.

[7] Kauko, T; Hooimeijer, P; Hakfoort, J, Capturing housing market segmentation: An alternative approach based on neural networkmodelling[J]. HOUSING STUDIES,2002,17: 875-894.

[8] DerrickS. Booneand Michelle Roehm, Evaluating the Appropriateness of Market Segmentation Solutions Using Artificial Neural Networks and the Membership Clustering Criterion[J]. Marketing Letters,2002: 317-333.

[9] Velu, C M,Kashwan, K R. Artificial Neural Network Based Data Mining Technique for Customer Classification for Market Forecasting [J]. International Journal of Advancements in Computing Technology,2015,71.

[10] Sun, JianshanWang, GangCheng, XusenFu Yelin, Mining affective text to improve social media item recommendation [J] Information Processing & Management. 2015, 51(4): 444-457.

[11] Oded Netzer, Ronen Feldman, Jacob Goldenberg and Moshe Fresko, Mine Your Own Business: Market-Structure Surveillance Through Text Mining [J]. Marketing Science, 2012, 31(3): 521-543.