PETROCHEMICAL PRODUCTION BIG DATA AND ITS FOUR TYPICAL APPLICATION PARADIGMS

Hu Shaolin*, Zhang Qinghua, Su Naiquan, and Li Xiwu

Guangdong University of Petrochemical Technology, Maoming, Guangdong, China

ABSTRACT

In recent years, the big data has attracted more and more attention. It can bring us more information and broader perspective to analyse and deal with problems than the conventional situation. However, so far, there is no widely acceptable and measurable definition for the term "big data". For example, what significant features a data set needs to have can be called big data, and how large a data set is can be called big data, and so on. Although the "5V" description widely used in textbooks has been tried to solve the above problems in many big data literatures, "5V" still has significant shortcomings and limitations, and is not suitable for completely describing big data problems in practical fields such as industrial production. Therefore, this paper creatively puts forward the new concept of data cloud and the data cloud-based "3M" descriptive definition of big data, which refers to a wide range of data sources (Multi-source), ultra-high dimensions (Multi-dimensional) and a long enough time span (Multi-spatiotemporal). Based on the 3M description of big data, this paper sets up four typical application paradigms for the production big data, analyses the typical application of four paradigms of big data, and lays the foundation for applications of big data from petrochemical industry.

KEYWORDS

Big Data, Paradigms, Industrial Big Data.

1. INTRODUCTION

In 1980, Alvin Toffler, a famous American futurist, first proposed the concept of big data in his book "The Third Wave". The concept of big data was first put forward in 2010 when nature launched the big data special issue. From then on, the big data as a term has been widely used in international academic circles and application fields.

Since the concept of big data was accepted by the Chinese in 2013, a big wave of big data has emerged across the country. Not only has the government incorporated big data into the national development plan and established a series of big data research institutions or big data centres [1], but also big data appears in people's daily life in a common language.

Over the years, although the concept of big data is more and more concerned by people in different fields and applied in various situations, some basic problems [2,3] about big data have not been solved yet, such as how to define the boundary between big data and conventional data, how to measure it, how to fully utilize the rich information of big data, etc.

Up to now, the big data definition widely adopted by almost everyone is the descriptive definition of "5V" that is the 5V characteristics of big data proposed by IBM: volume (large amount), velocity (high speed), variety (diversity), value (low value density), and veracity (authenticity). However, parts of these 5V are inappropriate or ambiguous [1,3]. For example, in a long-running

actual petrochemical plant, a single-dimensional long-term sampling data sequence collected by a sensor for quite a long time is not enough to constitute the so-called "big data" in the field of big data research, although the volume of this data series can be very large over time. For another example, the fourth "V" in 5V, that is, low value density, is a hard to understand feature for industrial big data. In the field of industrial production, such as petrochemical process, some sampling data may not be of great value to process analysis, but from another perspective, we will find a lot of valuable information, the key point of which depends on which perspective you look at the problem. Some literatures define big data as “big” data which is difficult to deal with by all the existing computer software[5-8]. Obviously, most of these definitions are descriptive and not rigorous enough, and difficult to distinguish between big data and conventional data.

In order to overcome the limitations of 5V definition of big data, and to accurately describe the characteristics of big data, section 2 will present a new set of characteristic definitions of big data from a new perspective, and explore the approach to measure the volume of big data. Based on the new definition of big data, section 3 will briefly describe several typical paradigms of big data applications. These typical paradigms are helpful for us to understand the actual use of big data and to grasp the essentials of how to implement these practical applications. At the end of this paper, several conclusions will be refined.

2. MEASURABLE DEFINITION FOR INDUSTRIAL BIG DATA

For the existing concept of big data, "big" is a fuzzy, imprecise adjective that is difficult to define its scope. The "5V" definition of big data does not strictly show which data set is big data, how to measure and judge whether a data set is big data. In order to overcome these limitations about the definition of big data[2,4], this section presents a new measurable 3M definition for industrial big data.

2.1. Description of Industrial Big Data

Generally, the manufacturing or production process may last for a period of time, for example, the production process of petrochemical production line lasts for several days, months, or even years. The production process may be repeated, and the products may be affected by various internal and external factors such as raw materials, environment, processing disturbance, etc. Therefore, the data related to petrochemical production process have many kinds, wide sources, various forms, different mechanisms and various uses. As shown in Figure 1.

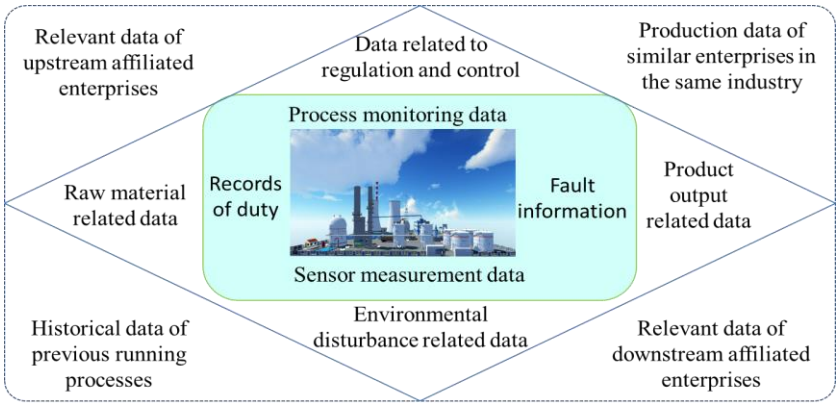

Figure 1. Three levels of data related to petrochemical production

All these data generally include at least six categories: historical data and current field data; operating mode data and sensing measurement data; raw material data and product related data; normal data production and abnormal production data; internal environment data and external environment data of production equipment; process control data and stochastic disturbance data, etc. In addition, it also involves the production data of related upstream and downstream enterprises. It can be seen that for a petrochemical industrial production enterprise, the associated data will certainly constitute a complex big data set from the multi-source data network.

All kinds of data mentioned above seem to be disordered and disordered, but they are not isolated and unrelated to each other, but linked together for the purpose. We believe that big data is a purposeful activity. Because of a specific purpose, all kinds of seemingly unrelated and scattered data are strongly or weakly linked together, and the so-called big data is formed.

Taking the big data of petrochemical industry shown in Figure 1 as an example, the reason why we connect all kinds of data (including historical data, data of other related enterprises, internal and external data of production equipment, and environment of production environment) that do not seem to be closely related to each other is that our activities are purposeful. Our purpose is to analyse and evaluate the production process of petrochemical enterprises, such as production capacity, pollution, production efficiency, safety and security, all of which are inseparable from these data.

All these data are like clouds floating over petrochemical enterprises, including data clouds coming from other enterprises, data clouds generated in the history of the enterprise and new data clouds continuously generated in the production process of the enterprise. In order to achieve a different purpose, one or several data clouds will be pulled from these data clouds on demand.

2.2. Features and Definition of the Industrial Big Data

For the industrial big data composed of pieces of data clouds, denoting the cloud label set as Γ, each piece of data cloud can be recorded as S_λ $(\lambda \in \Gamma)$

$$S_\lambda = \{\mathrm{F}(\omega) \mid \omega \in \Omega_1 \times \Omega_2 \times \Omega_3\} \quad (\lambda \in \Gamma) \tag{1}$$

where, the functional F is a map from triples $\Omega_1 \times \Omega_2 \times \Omega_3$ to multidimensional data space, Ω_1 is a collection set of data structures or data types, Ω_2 is a collection set of various sampling dimensions or measurement channels of sensor networks, Ω_3 is the set corresponding to the sampling time or data period.

The cluster composed of pieces of data clouds as shown in formula (1) can be represented graphically, as shown in Figure 2.

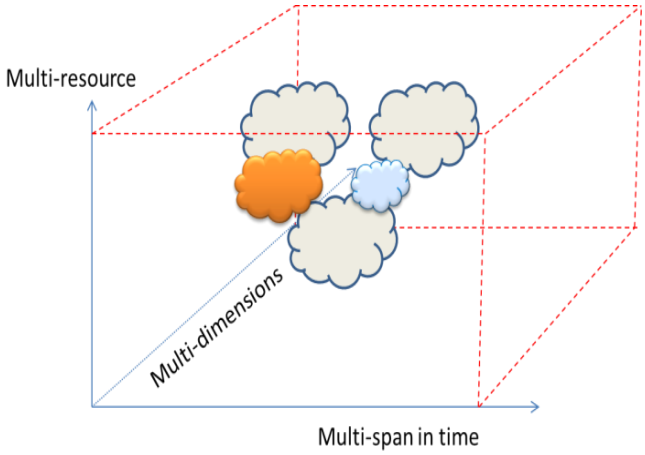

Figure 2. Big data cluster diagram

Figure 2 shows that the big data clusters are very much like clouds floating in the air one by one. Some of the data clouds may be separated from each other, and some of data clouds may overlap with other data clouds. And, it can be seen from Figure 2 that big data has "3M" features: multi-source, multi-dimension and multi-time and space.

(1) Multi-resource. In the field of industrial production and processing, data can come from different scenarios, different production periods, different production links, different objects, and different incentives. E.g In the petrochemical production process, data can come from the working conditions of petrochemical equipment, the production process of petrochemical products, the external environment of the petrochemical production process, the internal state of petroleum refining, the quality of petroleum raw materials, as well as personal factors, etc. All these data are important for the safety management of petrochemical processes. All these data come from different sources, just like a colourful cloud floating in the air, and they come together to form big data of petrochemical industry. In the petrochemical production management process, each cloud has its own rationality and necessity.

(2) Multi-dimension: When observing the real thing, it is usually viewed from different angles and different perspectives. Similarly, taking petrochemical production process as an example, a large number of sensor devices are set up in the production process. Each sensor is like a person's eye and observe a certain side of the chemical process to obtain information, which display the states of the process partially. Observation information obtained from different sides is brought together to form a set of high-dimensional data (even ultra-high-dimensional data). Although single-dimensional data is not enough to grasp the overall change, the multi-dimensional data flow can be used to completely describe the changing process of petrochemical production.

(3) Multi-span in time. Industrial production is an ongoing process. In a sense, actions performed at different times are repeated before or after other time points. The repetitive nature of process fragments is very useful for us to analyse process changes, judge working conditions, and diagnose abnormal working conditions. Data fragments at different periods or at different time points are also like colourful clouds floating in the data space. The advancement of data over time is an important feature of industrial big data and an important aspect of big data.

Based on the above description and analysis, this paper proposes a descriptive definition of industrial big data, which is a collection set B_S of data involving multi-source, high-dimensional, and spanning a long time span, as shown in equation (2):

$$B_S = \bigcup_{\lambda\in\Gamma} \mathrm{S}_\lambda = \bigcup_{\lambda\in\Gamma} \{\mathrm{F}(\omega) \mid \omega \in \Omega_1 \times \Omega_2 \times \Omega_3 \} \tag{2}$$

The volume $\|B_S\|$ of this collection set B_S is equal to the sum of the volumes of each data cloud, as shown in equation (3):

$$\|B_S\| = \sum_{\lambda\in\Gamma} \|\{\mathrm{F}(\omega) \mid \omega \in \Omega_1 \times \Omega_2 \times \Omega_3 \}\| \tag{3}$$

There may be different data structures and data forms in industrial big data set. For example, part of the data in the big data set may be structured, or part of the data may be semi-structured or unstructured, or part of the data may be numerical, or part of the data may be image, or part of the data may be time domain, or part of the data may be frequency domain. In other words, the size of different data points in the industrial big data set may be different. Due to the above reasons, the size of big data is not equal to the capacity or volume of big data set, but the volume $\|B_S\|$ of big data set can reflect the size of big data to a certain extent. Moreover, if and only if $\|B_S\|$ is large enough, or at least one piece of data cloud is large enough, this data set is suitable to be called big data set.

3. FOUR TYPICAL APPLICATION PARADIGMS

Generally, industrial big data clusters are often widely used in the following four kinds of occasions: fusion calculation so as to improve accuracy for calculations and to use all usable information for statistics inference; model correction so as to provide a more basis for prediction and support decision making; information compensation so as to bridge over gaps between fragment information. The role of industrial big data cluster is different in different application occasions [6-9]. Correspondingly, the paradigm is also different.

3.1. Paradigm of Fusion Calculation

If the data clouds overlap, overlapping data clouds can give measurement information of objects from different perspectives. Making full use of overlapping data clouds is helpful for us to improve the accuracy of calculation results.

Data fusion is one of the important ways of big data application. Intuitive understanding is that different data will bring different information. Combing together these data, all these different information can be helpful to eliminate errors and to correct prejudices. In this way, we get more and more accurate results and approximate correct inferences. In other words, data fusion technology is an information processing technology which is used to analyse various observations under certain criteria so as to complete the required decision-making and evaluation tasks. Data fusion has achieved amazing development in the past 10 years and has entered many different application areas.

Data fusion is a purposeful activity. The industrial big data provides us with richer and more perspective data information. If we fully integrate the data information of different sources in big data, it is very valuable to improve the accuracy of inference or grasp the whole picture of things more comprehensively.

There are quite a lot of approaches for big data fusion. For example, for data layer fusion, the more widely used methods include least squares adjustment, etc.; for information layer and

decision layer fusion, if all sheets of the floating data clouds $S_\lambda\ (\lambda \in \Gamma)$ are independent, we may use the Bayesian inference and stochastic decision making:

$$P(E \mid B_s) = \frac{\sum_{\lambda \in \Gamma} P(ES_\lambda)}{\sum_{\lambda \in \Gamma} P(S_\lambda)} = \sum_{\lambda \in \Gamma} \frac{P(ES_\lambda)}{\sum_{\lambda \in \Gamma} P(S_\lambda)} = \sum_{\lambda \in \Gamma} \frac{P(S_\lambda)}{\sum_{\lambda \in \Gamma} P(S_\lambda)} P(E \mid S_\lambda) \tag{4}$$

where E is the event to be inferred.

In the big data environment, not every piece of floating data cloud is bound to participate in the integration. The condition of industrial big data fusion is that the data cloud that can participate in the fusion must be consistent in time, space or object connotation. At least, if a piece of data cloud participates in the fusion, the data cloud and other data clouds involved in fusion processing must be able to overlap or partially overlap after time traversal or space conversion.

3.2. Paradigm of Information Compensation

The industrial big data is the aggregation of quite a large number of data from different sources. This kind of big data has a wide range of sources and temporal spatial characteristics, which is the advantage of big data, so we have the opportunity to deepen the understanding of objects from different perspectives.

Taking petrochemical big data-driven production process safety monitoring as an example, deep mining and modelling of historical data in industry big data can be used to understand the characteristic change of production process and judge the cause of current state and accurately predict its future change trend; the historical influence of environmental disturbance on the past state is helpful to judge the influence of environmental disturbance on the current state change, and to adjust the monitoring threshold of state safety. In addition, the input raw material data and output quality data can be used to analyse the internal working status of some key equipment such as cracking furnace.

There is a metaphor that conventional scale data allows us to see several aspects of things or objects, rather than overall impression, which is similar to six blind people touch an elephant, and each one touches only one side of the elephant. The multi-source heterogeneity of big data can just be used to fill the information gap that has not been touched, as shown in Figure 3.
Based on this consideration, the second typical application paradigm of big data is to fill and repair the information gaps. Specifically, a map Ψ is suitably constructed from the data cloud to the feature subspace:

$$\Psi : B_S \to \Theta \tag{5}$$

This map has the following properties: if the big data cloud sheet S_i contains the information of interest, then $\Psi(S_i) \subset \Theta$; otherwise, $\Psi(S_i) = \Phi$, an empty set. What we want to do is to find all the cloud sheets mapped to the non-empty sets from the big data cloud sheet.

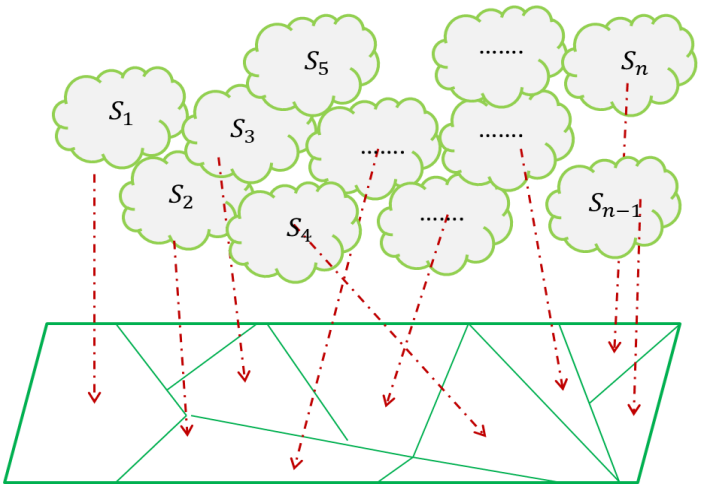

Figure 3. Mapping from big data to feature space

Obviously, if the union of the phase space of big data cloud can cover the whole feature space of interest, it is possible to realize the complete understanding of the feature of interest by using big data. Otherwise, even if the amount of big data is large enough, it is impossible to fully understand the changes in the characteristics of big data cloud.

This paradigm stated above is helpful for us to adopt a problem-driven approach to delete or reduce unnecessary data blocks to achieve big data compression.

3.3. Paradigm of Transfer learning and Knowledge Inheritance

There is an old Chinese saying that "stones from other mountains can be used to carve beautiful jade". The big data of petrochemical industry shown in Figure 1 not only has the data related to the production process of petrochemical plants concerned, but also has the data accumulation of production process of other similar plants. The knowledge contained in the mass production data from other similar enterprises can be used to improve and optimize the production process.

The development of artificial intelligence, especially the theory and model of machine learning and deep learning, provides a feasible technical way to use the knowledge accumulated by other similar enterprises to improve their production.

The combination of machine learning, big data and artificial intelligence is a measure to solve the difficult problem of big data application. This section describes a data modelling logic that combines big data with transfer learning[10,11], as shown in Figure 4.

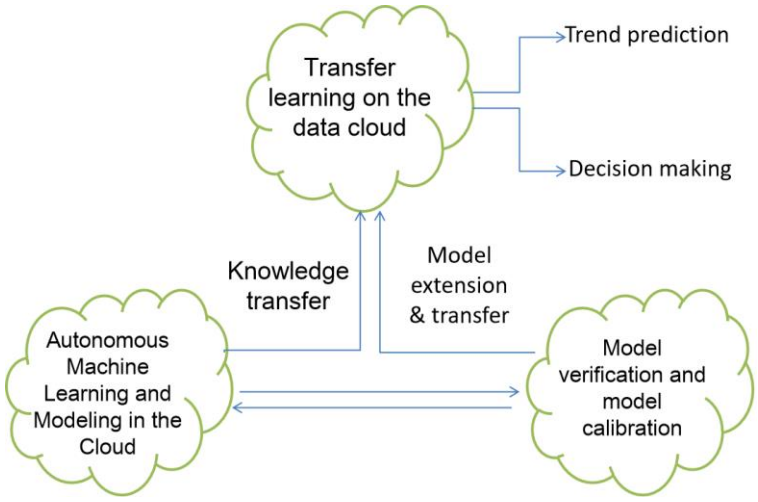

Figure 4. Prediction and decision based on transfer learning

Applying big data analysis technology to specific industrial big data, the above paradigm can ensure that the big data application process is understandable and interpretable. The advantage of the above paradigm based on big data transfer learning is that the learning process is completed in the cloud chip. Due to the relative consistency of the data structure in the cloud chip, the learning process is relatively simple and easy to implement; the data of other cloud chips is used to verify the learning model and perfection to ensure the inheritance and usability of knowledge and models.

3.4. Paradigm of Process Monitoring and Fault Prevent

The value of big data is that it can provide more data resources, information and knowledge for application, and so is industrial big data. The fourth paradigm focuses on the goal of ensuring the safe and stable operation of industrial system, and discusses how to give full play to the advantages of big data to better realize the monitoring, fault diagnosis and fault prevention of industrial production process, and improve the system safety.

Ensuring the safety of industrial production process system is a complex system engineering, which usually involves four aspects, including production process condition monitoring, situation awareness, fault diagnosis, fault prediction, fault disposal, health management and life extension. Each of the above aspects is an important research area, and various data-driven methods[11-14] have been proposed in recent years. Under the framework of big data, it is expected to form an integrated system security management and control system driven by big data by integrating or fusing different data-driven methods, as shown in Figure 5. This is an important way of big data application in the future, and can be extended to a wide range of fields, such as factory production, spacecraft operation and maintenance, urban traffic infrastructure security regulation and urban flood control and disaster prevention.

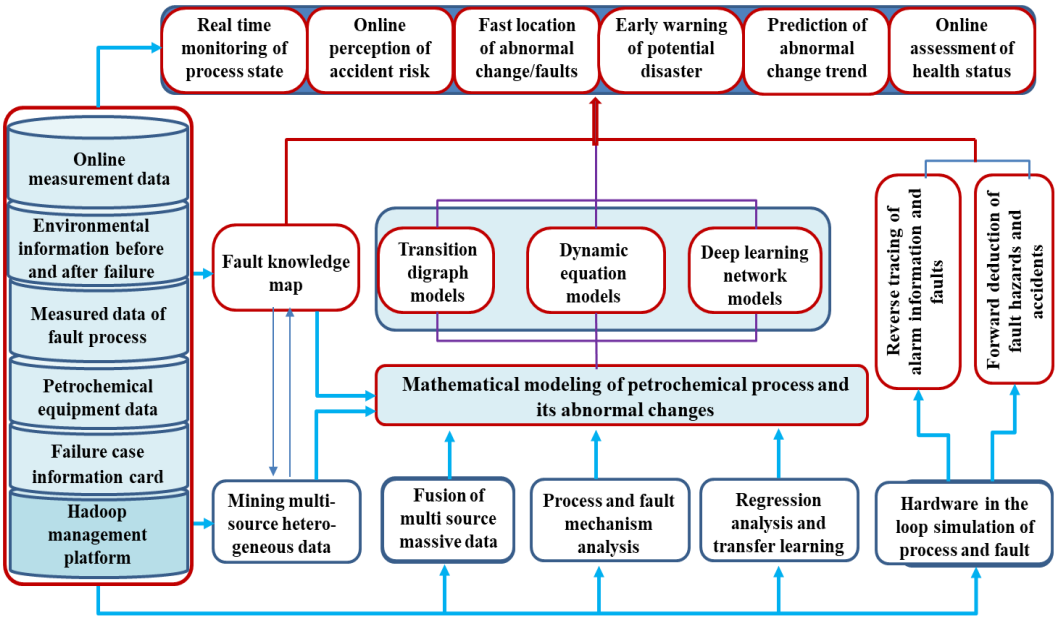

Figure 5. Monitoring, prognostics and early warning driven by industrial big data

Although the big data research includes big data acquisition, big data management, big data processing and big data application, Figure 5 focuses on the application of big data. Starting from the big data of petrochemical industry, through the mining, analysis, learning and extraction of knowledge in the big data, the construction of knowledge map and the data modelling of operation process are realized. By combining knowledge mapping, model and process simulation, the application system integrates six functions, including the real-time monitoring of process state, online perception of accident risk, fast location of abnormal change/faults, early warning of potential disaster, prediction of abnormal change trend, and online assessment of health status, etc.

4. CONCLUSIONS

In this paper, we intuitively put forward the visual description of industrial big data cloud and the clouds based 3M definition for industrial big data, and point out that the 3M definition not only well covers the three basic characteristics of industrial big data, including multi-source (extensive source), multi-dimension (complex structure) and multi-spatiotemporal (spanning different time and space), but also overcomes the limitations of the existing 5V definition in many aspects, such as unclear meaning and difficulty in quantifying the scale of big data. More importantly, the 3M definition of big data is a step forward in measuring the scale of industrial big data.

This paper puts forward a new and important view point: big data is a purpose oriented activity. Generally, a big data set is the aggregation of quite a large number of purpose oriented data. The reason why all kinds of seemingly disordered data are linked together must be because of some purpose. For the specific purpose, various seemingly unrelated and scattered data are strongly or weakly linked together.

On this basis, four typical paradigms of industrial big data processing and application are given, including multi-source information fusion of the industrial big data, information integration and compensation based on big data, inheritance and transfer learning knowledge contained in different data sources, and the practical integration paradigm based on industrial big data for petrochemical system safety.

These four typical application paradigms cover the basic form of big data processing and application in the industrial field, which is helpful to systematically grasp the application method of industrial big data and the effective approaches of demand-oriented application of big data to solve practical problems.

ACKNOWLEDGEMENTS

This paper is financially supported by the Nature Science Foundation of China (61973094, 61933013), the Maoming Nature Science Foundation (2020S004), and the Guangdong Basic and Applied Basic Research Fund Project (2020B1515310003).

REFERENCES

[1] Li Ning(2019).Artificial Intelligence Paradigm in Big Data Era. China Computer & Communication, 8:104-105.
[2] Jarosław W, Jarosław J, Paweł Z(2019), Generalised framework for Multi-criteria Method Selection. Omega,86:107–124.
[3] Manyika J, Chui M, Brown B, Bughin J.,et al(2017). Big data: The Next Frontier for Innovation, Competition, and productivity. McKinsey Global Institute
[4] Abbass H. A, Leu G,Merrick K(2016).A Review of Theoretical and Practical Challenges of Trusted Autonomy in Big Data. Theoretical Foundations for Big Data Applications: Challenges and Opportunities.
[5] Ammu N, Irfanuddin M. Big data challenges. International Journal of Advanced Trends in Computer Science and Engineering, 2013,2(1), 613-615
[6] Rabhi L, Falin N, Afraites A,et al(2019).Procedia Computer Science,16th International Conf. on Mobile Systems and Pervasive Computing, v155, pp:599-605.
[7] Mukherjee S, Shaw R(2016). Big Data-Concepts, Applications, Challenges and Future Scope. Wikipedia, https://en.wikipedia.org/wiki/Google_Cloud_Platform
[8] Tanya Garg;Surbhi Khullar(2020). Big Data Analytics: Applications, Challenges & Future Directions. 8th International Confer on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO). pp:923-928,4-5 June 2020, Noida, India.
[9] Nojood Aljehane(2020).Big Data Analytics: Challenges and Opportunities. International Confer on Computing and Information Technology (ICCIT-1441), pp:1-4, 9-10 Sept. 2020, Tabuk, Saudi Arabia.
[10] Ma Z, Yang L, Zhang, Q(2021). Support Multimode Tensor Machine for Multiple Classification on Industrial Big Data. IEEE Transactions on Industrial Informatics,17(5):3382-3390.
[11] Sakineti S; Prabhu C (2018). Protagonist of Big Data and Predictive Analytics using Data Analytics. International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS), pp:276-279,21-22 Dec. 2018, Belgaum,India.
[12] Zhou X, Hu Y, Liang W,et al(2021). Variational LSTM Enhanced Anomaly Detection for Industrial Big Data. IEEE Transactions on Industrial Informatics,17(5): 3469-3477.
[13] Sfaxi L,Ben A,Mohamed M(2020), DECIDE: An Agile Event-and-Data Driven Design Methodology for Decisional Big Data Projects. Knowledge Engineering, volume 130, DOI: 10.1016/j.datak.2020.101862.
[14] Battistelli G, Tesi, P(2021). Classification for Dynamical Systems: Model-Based and Data-Driven Approaches.IEEE Transactions on Automatic Control, 66(4):1741-1748.

AUTHOR

Hu Shaolin received the Ph.D. degree in system engineering from Xi'an Jiaotong University in 2000.From 2006 to 2007, he was a visiting scholar with the Royal Institute of Swedish. He is the author of five books, more than 200 articles, and more than 70 inventions. His research interests include system safety, process monitoring, fault diagnosis, big data processing and data clustering, exploratory data analysis, navigation and control.

Mr. Hu is the senior member of the Chinese Automation Association (CAA) and senior member of the Chinese Association of Artificial Intelligence (CAAI). He was a recipient of Swedish Institute Visiting Scholar Award in 2006 and winner of China Aerospace Fund Award in 2011.