Research on Interactive Modeling and Intelligent Analysis Platform Based on Railway Big Data

Xiaodong Zhang¹, Ping Li², Wei Wu¹ and Qingmeng Liu¹

¹Institute of Computing Technology, China Academy of Railway Sciences, China
²China Academy of Railway Sciences, China

*Corresponding author. Email: xdcheung@126.com

Abstract. With completing of the Railway Data Service Platform (RDSP), the railway industry has a better big data foundation. The railway industry faces difficulties in applying and promoting artificial intelligence. Corroborating the batch data and real-time data, an intelligent corroborative analysis and learning method is constructed. Based on the storage and computing abilities of RDSP, the general artificial intelligence algorithms are embedded. Three types of railway data analysis model units are reconstructed. Try to realize the visualization and interaction of railway business modeling, which can simplify the data analysis. Based on the knowledge structure level of the railway business staff, the data analysis interaction types are designed to different models. Design and set up the prototype of Interactive Intelligent Analysis Platform based on Railway Big Data. The prototype platform improves the capacities of automatic learning from data knowledge and ensures the validity and timeliness of the data analysis.

Keywords: railway, big data, artificial intelligence, interactive modeling, data fusion.

1. Introduction
The emergence of information technology and intelligent technology has brought opportunities for China railway development in the new era. The integration of intelligent technology and railway business will promote the intelligent railway development. The core features of intelligent railway are comprehensive perception, allover interconnection, fusion processing, active learning and scientific decision-making [1]. With the establishment of China Railway Main Data Center, a huge amount of data has been collected to form the railway big data asset lake. The railway big data can provide data support for the three characteristics of intelligent railway, comprehensive perception, all over interconnection and fusion processing. On the other hand, active learning and scientific decision-making need to base the intelligent analysis technology to deeply excavate the great value contained in massive data assets. This can provide scientific and reliable support for railway production.

For the new era, big data related technology and data acquisition and transmission technology develop rapidly. Big data analysis has replaced the original data analysis. It is one of the core technologies to realize railway intelligence [2, 3]. Through allover perception of railway mobile equipment, fixed infrastructure, natural environment and transport service status, the RDSP gets and stores much structural, semi-structured and unstructured data. These data have direct or potential value
to train operation failure warning, condition-based maintenance, precision marketing and so on. How to mine and learn the important value implied in these data to guide the railway production is the most concerned issue for the railway department.

The promotion of RDSP in the whole China railway has been carried out one after another. Each railway bureau corporation has carried on the comprehensive management about the big data of the railway bureau. All the management is according to the *Big Data Application Implementation Plan of National Railway Corporation, China*. At the same time, the problem of how to mine the value of data, that is, how to use it, has become increasingly prominent. Intelligent mining and learning the value hidden in the massive data is important to guide the railway production. It has become a key problem to improve the use of railway big data. It is difficult to transform a railway application scene into mathematical model or a computer model. This involves complex tasks such as data distribution evaluation, data independence test, data model comparison and selection, data model evaluation, specific coding development, model out-put, timing task formulation and so on. The knowledge reserve and skill of railway staff are high. While the education level of railway staff is uneven. Then it is difficult for big data and artificial intelligence to develop and apply in the railway industry. Therefore, it is necessary to carry out the research on intelligent analysis methods. The method should be high efficiency, reliability, flexibility and targeted under the condition of railway big data.

2. **Intelligent collaborative fusion method based on multi-temporal railway big data**

Many information management systems have been set up to serve various railway fields. These systems make up the source of railway big data. The total amount of China railway data is more than 10PB, and the daily growth is over 1TB. The rapid growth in the scale and variety of railway data shows that China railway has entered the era of big data. China Railway Corporation settled the Main Data Center in Wuqing District, Tianjin. The Railway Data Service Platform (RDSP) stores all the data.

There are many supervised learning models in the railway business. Supervised learning learns or builds a model from training data and infers new examples based on this model [4]. Railway business has dynamic massive data. Thus, supervised learning in the traditional sense is not perfect enough. The disturbance of abnormal data and the expansion of data dimensions will cause analysis dilemmas with the existing business models, as shown in Figure 1.

![Figure 1. New characteristic content appearing in the incremental data.](image)

An intelligent collaborative fusion method based on multi-temporal railway big data focus on the relationship between the deposited historical data and real-time data. It coordinates the historical data with the real-time data. The research mainly carried out from three main technologies: batch data processing technology, real-time data stream processing technology and collaborative evolution algorithms, as shown in Figure 2.

![Figure 2. Schematic diagram of the intelligent collaborative fusion method.](image)
The batch data is history data, which is always important and nonurgent. Batch data processing technology uses the history data to build up the model. It helps to form railway business models and provides history solutions for railway production. Streaming data is the instant data produced by the production system, which is always important and urgent. When the business system has real-time data, the platform recommended to use the existing railway professional model to analyze. That can quickly respond to railway production needs. Intelligent collaborative fusion method can fuse batch and streaming data. Satisfying the time trigger condition or event trigger condition, the platform should judge that whether the new feature appear in the history data and whether there is abnormal data disturbance. If the situation met the conditions, the platform will adjust or remodel the railway professional model. This method ensures the timeliness and accuracy of railway professional model.

3. Technical structure of interactive intelligent analysis platform based on railway big data

Interactive Intelligent Analysis Platform based on Railway Big Data (IIAP) is a new generation of intelligent data mining and analysis system serving the railway. The overall technical architecture of IIAP is shown in Figure 3.

There are 9 parts and 5 layers in the architecture. The bottom layer of the platform consists of Data Management and Data Resource Interface Access. The bottom layer uses the currently popular Hadoop framework for data storage and a Spark framework for distributed parallel computing of the data. In the part of Data Management, structured and unstructured data are stored in different media, for example, Hive, HDFS, relational databases. This part also sets up the data resource catalog, meta-data and master data. The platform interfaces transfer the data by JDBC, Spark SQL, etc. The middle layer is composed of Data Computing and Processing, Basic Application Service, Data Sharing Service and Data Application Support. In the part of Data Computing and Processing, all models and algorithms use Spark JobServer to schedule tasks. In the part of Basic Application Service, the platform encapsulates open source machine learning algorithms and statistical algorithms based on Spark MLlib and Spark R by components. Data Sharing Service shares data, calculations and visualizations and Data Application Support uses JavaScript and HTML5 to implement visual interaction modeling of the modular algorithms. The display layer of IIAP uses the ZK framework to response the request of data, modeling, computing, application, scheduling and other services, which consists of User Session. The user authentication and authorization are also considered. Security System ensures that is away from external risks. Service and Standard Guarantee System provide external service standards to converge services between the different one. The platform as a whole has a high-performance memory computing level. It carries out deep customization of models and algorithms for railway business.
Figure 3. Technical Structure of IIAP

The platform integrates 152 mainstream statistical methods and machine learning methods in 5 major categories. The statistical methods are based on Spark R. Machine learning methods are based on Spark MLlib. Among them, data preprocessing algorithms include 36 algorithms such as data fusion and time processing. Machine learning algorithms include 26 algorithms such as hybrid clustering and association recommendation. Statistical algorithms include 73 algorithms such as dispersion test and correlation analysis. Model evaluation algorithms include 17 algorithms such as classification evaluation and regression evaluation. In addition, the platform also integrates Jupiter development tools for users to model.

IIAP combines the railway production needs. The railway professional AI model is formed from three parts. That is space-time distribution relation algorithm, index correlation algorithm and behavior analysis algorithm. It simplifies the modeling, reduces the use threshold, and provides intellectual support for mining the hidden value of railway big data.

4. Function design of interactive intelligent analysis platform based on railway big data
Interactive Intelligent Analysis Platform based on Railway Big Data (IIAP) includes 7 major function modules, as shown in Figure 4. Data & File Module manages the different type data. Machine Learning & Mining Module contains all the data mining function based on Spark MLlib. Statistical Analysis Module contains all the methods based on Spark R. Railway Professional Model Module contains the
railway custom analysis method. BI Report Module uses the result data to make visual charts and tables. Case Management Module stores the users’ cases. System Management Module contains the system basic functions.

**Figure 4.** Function modules of IIAP.

Combined with railway major classification, different data analysis models and interactive processes are carried out for the production commanders, measurement statisticians, and data analysts in IIAP. It targets to solve the pain points of railway big data and artificial intelligence applications. In addition, it expands the scope of railway artificial intelligence.

After the railway daily production data enters IIAP, the platform provides multiple railway data analysis services and cleaning system boundaries for different users, as shown in Figure 5. We call the staff from the transportation bureau, construction management department, vehicle management department, et al the production commanders. They have many people and wide distribution, and they have a strong demand for daily production instruction tables and sheets. This kind of demand is generally realized through the statistical analysis function of the platform, including the common statistical units in SPSS. Thereby, instruction data is generated. As shown in the orange area of Figure 5, their work costs have been reduced and their work efficiency has been improved. We call the staff from the finance department, railway audit department, et al the measurement statisticians. Their education level is relatively high. They need the visualization function of IIAP to analyze the railway big data simply, and the friendly display charts could be formed. The data accumulated in the areas of concern are analyzed periodically in terms of day, week, month, quarter, year and so on. As shown in the blue area of Figure 5, IIAP improves the analysis effectiveness and the measurement statisticians can find more obvious internal rules. The data analysts refer to the professional data analysts at the National Railway Corporation and Railway Bureau Corporation. They have fewer people with certain data analysis methods relevant railway knowledge. IIAP has the characteristics of rich development languages, complete intelligent algorithms, flexible and free collocation, convenient data dimensionality descent, independent development and combination, high-speed parallel operating and so on. As shown in the red area of Figure 5, IIAP can help to improve the potential of data analysis and mining in an all-round way and provide convenience for data mining.

**Figure 5.** Service objects and boundaries of IIAP.
5. Conclusions
In the next 3 to 5 years, with the maturity of artificial intelligence technology, the intelligent high-speed railway innovation practice will continue to break through. The scale of big data in the railway field will also increase exponentially. Intelligent analysis needs of railway will become more and more urgent. Turning data into value will become the most important task in the future. It directly affects the improvement of the intelligence level of the railway. The research can complement the relevant changes of the market in advance and achieve improvements in intelligent analysis.

Acknowledgement
This research was financially supported by the Institute of Computing Technology, CARS, studied by Research and Application Innovation Center for Big Data Technology in Railway. The project number is DZYF18-22.

References
[1] Shao Sai, Li Ping, Zhen Jinzi, Study on the Evaluation Index System of China Intelligent High-speed Railway Based on Technical Characteristic, IEEE Intelligent Transportation Systems Conference. 2019.
[2] Zhang Xiaodong, Ma Xiaoning, Li Ping, Wu Wei, Research on the Application and Development of Artificial Intelligence in China Railway, China Railway, 2019 (11), pp. 32-38.
[3] Wang Tongjun, Research and Application of China Intelligent High-Speed Railway Architecture, Journal of The China Railway Society, 2019 41 (11), pp. 1-9.
[4] Hefny A, Downey C, Gordon G, J, Supervised Learning for Dynamical System Learning, International Conference on Neural Information Processing Systems. MIT Press, 2015.