Constructing Knowledge Graph for Prognostics and Health Management of On-board Train Control System Based on Big Data and XGBoost

Jiang Liu[1,3](✉), Bai-gen Cai[2,3], Zhong-bin Guo[1], and Xiao-lin Zhao[1]

[1] School of Electronic and Information Engineering,
Beijing Jiaotong University, Beijing 100044, China
jiangliu@bjtu.edu.cn

[2] School of Computer and Information Technology,
Beijing Jiaotong University, Beijing 100044, China

[3] Beijing Engineering Research Center of EMC and GNSS Technology
for Rail Transportation, Beijing 100044, China

Abstract. Train control system plays a significant role in safe and efficient operation of the railway transport system. In order to enhance the system capability and cost efficiency from a full life cycle perspective, the establishment of a Condition-based Maintenance (CBM) scheme will be beneficial to both the currently in use and next generation train control systems. Due to the complexity of the fault mechanism of on-board train control system, a data-driven method is of great necessity to enable the Prognostics and Health Management (PHM) for the equipments in field operation. In this paper, we propose a big data platform to realize the storage, management and processing of historical field data from on-board train control equipments. Specifically, we focus on constructing the Knowledge Graph (KG) of typical faults. The Extreme Gradient Boosting (XGBoost) method is adopted to build big-data-enabled training models, which reveal the distribution of the feature importance and quantitatively evaluate the fault correlation of all related features. The presented scheme is demonstrated by a big data platform with incremental field data sets from railway operation process. Case study results show that this scheme can derive knowledge graph of specific system fault and reveal the relevance of features effectively.

Keywords: Knowledge graph · Prognostics and health management · Train control system · Big data · Machine learning · Extreme gradient boosting

1 Introduction

As the economic and social development all over the world, the demand for railway passenger and freight transport is expected to grow in the future. It is a common sense that safety is the heart of all the operation-related activities in railway transportation. As the core safety system controlling the movements of all moving trains, the Train

Z. Deze et al. (Eds.): BDTA 2020/WiCON 2020, LNICST 371, pp. 3–16, 2021.
https://doi.org/10.1007/978-3-030-72802-1_1

Control System (TCS) is responsible of setting up non-conflicting and safe routes for trains, defining safe limits of movement, and transmitting instructions or commands to train drivers. The high-speed railway is experiencing a rapid development in many countries. Presently, the railway industry has taken a great effort and is focused on the exploitation of advanced ABCDE (Artificial intelligence, Block chain, Cloud computing, big Data, Everything is connected) technologies in the new generation train control systems. Specifically, the big data technology has been applied in many aspects of railway operation and maintenance, including railway condition monitoring [1], train delay analysis and prediction [2], passenger route choice and demand forecasting [3], railway event and accident analysis [4], and optimization of time scheduling [5]. The integration of improved information conditions and advanced information processing technologies makes it possible of achieving a flexible, real-time, intelligent, integrated and fully automated railway operation and management system.

In the operation of the high-speed railway, fault diagnosis and equipment maintenance for train control systems play a significant role in ensuring safety and efficiency of the trains. In order to effectively extend the system life cycle, several new maintenance concepts have been proposed and utilized in many industrial branches. Particularly, the Prognostics and Health Management (PHM) technique, which aims to carry out fault detection, condition assessment and failure prediction, has been successfully implemented in the railway domain [6, 7]. However, conventional corrective maintenance and time-based maintenance strategies are still adopted in the maintenance and management of high-speed train control systems in China. Due to the limitations and conservative characteristics of the in-use maintenance rulebook, there is still space for improving the utilization of Remaining Useful Life (RUL) of train control systems, which is of great importance to optimize the cost efficiency of maintenance activities from a full-life-cycle perspective. The development of the novel train control systems requires the concentration on advanced methodologies and intelligent technologies to enable the Condition-based Maintenance (CBM), which is regarded as the key issue to cope with the increasing complexity of influencing factors on the competitiveness.

For the on-board train control system, CBM can be realized based on the enhancement of state monitoring conditions to the whole system and its components. However, it is difficult to add extra monitoring sensors or units, which are not dedicated to train control functions, to precisely collect the expected state variables. Consideration to the safety-critical characteristic and risk controlling against the system complexity makes it difficult to collect required system performance metrics directly for generating optimized maintenance decisions. Fortunately, the on-board train control system is designed with a capability of recording data logs in real-time during the field operation. In these log files, a number of status and state data fields at both the component level and system level are recorded along with fault/failure flags with respect to corresponding predefined fault modes. The accumulative data logs enable the opportunity to investigate the fault causes and the development characteristics under a rich information condition, which encourages us to establish a specific decision support system for the condition-based maintenance of the on-board train control equipments under the PHM framework through a big data-based approach.

In this paper, we mainly focus on the construction of the Knowledge Graph (KG) of specific on-board train control systems, which is an important foundation to realize the quantitative prediction of fault probability and risks to perform decision making in active maintenance. The characteristics of fault modes adopted in field operations are analyzed. Architecture of the big data platform and the XGBoost-enabled knowledge graph generation method by model training are introduced. Case study results with historical data sets from practical on-board train control equipments are given to show the capability of the presented solution.

The rest of this paper is organized as follows. Section 2 gives a global description to the architecture of the big data-based platform. In Sect. 3, the knowledge graph construction method using the XGBoost algorithm is introduced in detail. Section 4 depicts the realization of the platform and reports the results in the case study. Finally, conclusions and future plans are presented in Sect. 5.

2 Architecture of Big Data-Based Platform

The on-board train control equipment is a core part of Chinese Train Control System (CTCS), which is designed for safety assurance of the railway system by preventing the trains from over-speeding during the tracing between a train and its leading target train. Enormous effort has been devoted to enhance the safety protection capability of the whole system by advanced redundant architectures, fail-safe logics and interfaces, complete system testing and the verification specifications. The status monitoring and event recording of the on-board equipment provides direct information sources for the operators to analyze the operation state and find out the causes of the recent malfunction(s). Taking the 200H on-board equipment for high-speed railway as an example, the system developers have concerned the data logging function for a fault diagnostic purpose by using a PCMCIA card in the Data Record Unit (DRU). The maintenance staff will download the log files within the PCMCIA card when the train finished the planned operations. By using specific software tools, it is usually easy and intuitive to review and inspect the practical running status of equipment, actions of the Automatic Train Protection (ATP) unit, manual operation activities of the driver and the reported fault events. Through association analysis with data fields, curves and descriptions of the tools, the users can achieve fault diagnosis and maintenance determination according to the specifications and technical experience. However, the current event-driven maintenance mode only considers the local effect and possible responses to a specific fault report, and that is not sufficient to achieve a global system-level coverage to the equipment's whole life cycle. Under this circumstance, we proposed and developed a big data platform aiming at enabling a data-driven framework for the condition-based maintenance. By utilizing all the historical operation data, the association knowledge and occurrence regularity of different faults for the on-board train control equipment can be effectively obtained, which is of great value for the realization of the prognostics and health management over the current operation management rules.

Figure 1 shows the architecture of the big data-based platform for data management and CBM decision assistance of the on-board train control equipment. It can be found that the platform consists of four layers, which will be introduced as follows.

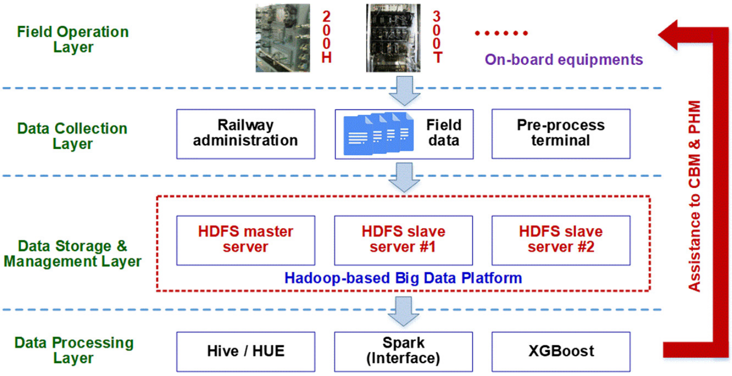

Fig. 1. Architecture of the big data-based platform.

2.1 Field Operation Layer

In this layer, on-board train control equipments record data logs to PCMCIA cards during the filed operation. Both the 200H and 300T on-board equipments are involved in the current version. System specifications and data extraction tools are involved in this front stage to collect useful data from the log files. With similar procedures, this platform is able to be compatible with other types of on-board equipments. Data acquisition based on this layer provides the foundation to reflect the equipment status in the time domain. At the same time, the derived knowledge graph and fault probability prediction results will provide feedback to this layer to affect the application of these on-board train control equipments in practical operation.

2.2 Data Collection Layer

The data collection layer is in charge of extracting and transforming raw data sets, which are obtained from the railway administration periodically. Specific tools and program scripts have been developed for automatic pre-processing, including the data classification, format transformation, code conversion and data cleaning. After these procedures, the characteristics corresponding to the fault labels, which are defined in the data logging tools according to the design of specific on-board equipments, would be prepared for incremental updating of the data storage and the following data processing operations.

2.3 Data Storage and Management Layer

This layer realizes the storage and management of the accumulative data collected from the equipments. The Hadoop Distributed File System (HDFS) [8, 9], which is an effective solution designed to run on the commodity hardware and has been proved suitable for the applications based on very large data sets, is adopted as the data storage framework. Three high level servers have been involved in this platform under a master/slave architecture. One server is configured as the NameNode to provide the service of managing the directory tree of all files and metadata about files and directories. The other two servers act as DataNodes in this Hadoop cluster to store the actual data files. This architecture ensures a flexible system extension capability corresponding to the increasing requirement and data quantity.

2.4 Data Processing Layer

The data processing layer achieves the exploration of the value from the field data files. Firstly, Hive is utilized as a data querying engine for the large-scale data stored in Hadoop. Through specific Python scripts, querying, sieving and statistical analysis can be realized according to the demand of the operators for equipment maintenance. The HUE (Hadoop User Experience) interface enables the users to access all the data files stored in HDFS and Hive-based data processing results. Secondly, Spark is integrated into the platform due to its advantages in iterative computing and accessibility to a set of machine learning APIs. Furthermore, the interface to the Extreme Gradient Boosting (XGBoost) models makes it possible to integrate the advanced algorithms with an enhanced computational efficiency level. Thirdly, XGBoost-based machine learning system is introduced for tree boosting to derive the feature importance scores for generating the knowledge graph under the Spark framework. Results of this layer finally achieve the goal of data-driven knowledge regularization for the different fault modes against the complicated and difficult mechanism modeling schemes.

3 Knowledge Graph Generation Method

Using a large amount of field data, it is difficult to clearly describe the mechanism of different faults or failures using analytical models that fit the data significantly better. From a data-driven perspective, the big data condition enables a different way to build knowledge models based on the long term state monitoring to the on-board train control equipment. The knowledge graph explores the potentials of heterogeneous data to reveal the relationship between a specific fault and all the related features. It provides a knowledge base of graph, where nodes indicate the entities and edges represent the relationship [10]. Through data training-based modeling, the derived KG is capable of identifying the contribution rate of each feature to a certain fault mode, and predicting the risks of the faults in future when the new-coming features are obtained. The most critical issue is the effectiveness of the knowledge graph. From the big data platform, we can extract sufficient training samples covering a long time period to ensure a rich information condition for model training. Thus, the adopted modeling solution will be the core issue that affects the performance of the derived KG and the trustworthiness of the KG-based prediction. In this paper, the XGBoost algorithm is adopted to build the knowledge graph based on big data. The identification of features and labels and the adopted algorithm are introduced as follows.

3.1 System Fault-Related Features

For the data logging tool of the 200H on-board equipments, the reported faults mainly attribute to six broad categories, including the controlling fault information, FSC error, SBUS error, DRU fault information, A/B system inconsistency and the STM fault information. Furthermore, each of the category corresponds to several sub-categories that describe the failures or abnormal states of certain components or units of an on-board equipment. Taking SBUS error as an example, there are seven sub-categories covering the information from the OPE, BUF and FSC. The fault status information is extracted as the labels to enable the model training.

Except the fault label information, there are still 121 data fields for a piece of log record. The data items indicate the time information, identity information, train operation condition, running state information, action status of key components, Driver Machine Interface (DMI) data and system parameters. A brief summary to these data entities for each class in one piece of record is given as follows.

(1) Time information
This class contains the data entities corresponding to the time instant when the data was recorded. The typical entities belong to this class include year, month, date, hour, minute, second, millisecond, etc.

(2) Identity information
The identity information indicates train unit number, active driver ID, train ID, etc. This information illustrates the identity of the train, on-board equipment and the driver to identify the target labels corresponding to each fault event.

(3) Train operation condition
The operation conditions resulted by the train control system describe the authority and moving space that the train has to follow. There are a series of condition-related data fields, e.g. the current track circuit length, next track circuit length, EBP (Emergency Brake Profile) speed, NBP (Normal Brake Profile) speed and the LMA (Limit of Movement Authority).

(4) Running state information
This class describes the in-trip running states of the target train, including the track location, actual speed, control speed, acceleration, accumulative running distance, etc.

(5) Action status of key components
The actions and operation status of specific components of the on-board train control equipment are recorded in real-time. This class includes action mode, brake order (VC1/VC2), EB brake indication, B7N brake indication, LSI information (‘A’ system /‘B’ system), online/inactive track circuit information, DRU information, OPE state, FSC state, STM state information, LKJ information, etc.

(6) DMI data
The DMI-related information reflects the operation-related information provided to the drivers for specific train control activities, e.g. DMI text information, DMI brake alert time, DMI target speed, DMI target distance, DMI switch category.

(7) System parameters
This class corresponds to parameters and configurations for the train and on-board equipment, e.g. the wheel diameter, train type, equipment/manual priority, DRU FSC ROM version, message class code.

Before the data sets are utilized in model training, all the features and labels represented by specific data entities have to be formatted in the data pre-processing stage, which enhances the storage efficiency and accessibility to the modeling algorithms.

3.2 XGBoost-Based Model Training

The XGBoost method, which was proposed in 2016 [11], is one effective supervised machine learning algorithm to realize a scalable tree boosting system. It has successfully attracted much attention due to its outstanding efficiency and high prediction accuracy in solving a number of practical problems [12–14]. Given a training dataset $\boldsymbol{D} = \{(\boldsymbol{x}_i, y_i)\}$ with n samples, $i = 1, 2, \cdots, n$, where $\boldsymbol{x}_i \in \mathbf{R}^m$ represents the variable with m features and y_i denotes the corresponding label, a tree ensemble model predicts the dependent variable $\hat{y}_i$ using the following model

$$\hat{y}_i = \Phi(\boldsymbol{x}_i) = \sum_{k=1}^{K} f_k(\boldsymbol{x}_i) \tag{1}$$

where $f_k \in \mathbf{F}$ denotes an regression tree with leaf scores, and $\mathbf{F}$ represents the space of trees as $\mathbf{F} = \{f(y) = \omega_{h(y)}\}$ with the leaf node $h(y)$ of the yth sample and the leaf score $\omega_{h(y)}$.

For the tth iteration, the prediction results can be

$$\hat{y}_i^t = \hat{y}_i^{t-1} + f_t(\boldsymbol{x}_i) \tag{2}$$

Thus, the objective function of the model can be written as the following form

$$J(f_t) = \sum_{i=1}^{n} L(y_i, \hat{y}_i^t) + \Gamma(f_t) = \sum_{i=1}^{n} L(y_i, \hat{y}_i^{t-1} + f_t(\boldsymbol{x}_i)) + \Gamma(f_t) \tag{3}$$

where $L(*)$ represents the loss function, $\Gamma(*)$ indicates the complexity of the model, and it can be defined using the number of leaf nodes T and score ω as follows

$$\Gamma(f_t) = \gamma T + \frac{\lambda}{2} \sum_{j=1}^{T} \omega_j^2 \tag{4}$$

By simplification of $\Gamma(f_t)$ using the second-order Taylor expansion, the objective function can be re-written as

$$J(f_t) = \sum_{i=1}^{n} \left[g_i \omega_{h(\mathbf{x}_i)} + \frac{1}{2} q_i \omega_{h(\mathbf{x}_i)}^2 \right] + \gamma T + \frac{\lambda}{2} \sum_{j=1}^{T} \omega_j^2 \tag{5}$$

$$g_i = \frac{\partial L(y_i, \hat{y}_i^{t-1})}{\partial \hat{y}_i^{t-1}} \tag{6}$$

$$q_i = \frac{\partial^2 L(y_i, \hat{y}_i^{t-1})}{\partial \hat{y}_i^{t-1}} \tag{7}$$

It can be optimized and the optimal solution can be represented by the optimal value of ω_j and the corresponding value of $\Gamma(f_t)$ as

$$\omega_j^* = -\frac{\sum_{i\in \boldsymbol{I}_j} g_i}{\sum_{i\in \boldsymbol{I}_j} q_i + \lambda} \tag{8}$$

$$J(f_t) = -\frac{1}{2}\sum_{j=1}^{T}\frac{\left(\sum_{i\in \boldsymbol{I}_j} g_i\right)^2}{\sum_{i\in \boldsymbol{I}_j} q_i + \lambda} + \gamma T \tag{9}$$

By training the XGBoost model, a series of decision trees can be obtained and used in classification with specific labels as described in the former section. Utilization of the second order Taylor expansion to a loss function as (4) and normalization against the over-fitting make it different from conventional methods like Gradient Boosting Decision Tree (GBDT) solution [15]. Details of this algorithm can be found in [11].

3.3 Generation of Knowledge Graph

Through the model training based on data samples with all related features and labels extracted from the big data platform, a set of trees can be obtained to reflect the importance of each feature. The derived model describes the relationship between features and the labels representing specific fault modes. It is obvious that the involved features contribute differently to a certain fault mode. The value of feature importance provides us a reference to quantitatively evaluate the content of contribution by each feature to a fault label.

By utilizing the feature importance values, the knowledge graph corresponding to a specific fault is represented by a set of triples $\{Node(fault), Link, Node(feature)\}$. As shown in Fig. 2, the structure of a local KG corresponding to a specific fault mode is based on a central node, m surrounding nodes and their links. $Node(fault)$ denotes the local center entity node indicating a specific fault mode from the fault label base. $Node(feature)$ represents the distributed graph node with respect to each of the features mentioned in Sect. 3.1. $Link$ in a triple formally illustrates a line connects $Node(fault)$ and $Node(feature)$, and it reveals the relationship between a feature and the fault mode quantitatively by the derived feature importance. It should be noticed that this example in Fig. 2 only shows a local area of the whole graph, which means there are more triples covering other central nodes $\{Node(fault)\}$ for different faults connecting all the involved m neighborhood feature nodes $\{Node(feature)\}$.

It has to be noticed that the knowledge graph established with the big data platform and the presented procedures is a typical data-driven solution. That means knowledge of the target fault modes reflected by the KG does not concerns the physical mechanism of the on-board train control system and the evolution rules of faults. It just explores the capabilities of huge historical data sets to build models that could reach a determination to the future health status and maintenance decisions to a target on-board train control equipment. This solution would not completely replace conventional mechanism-based models or analytical models, but enables a new path to reveal the feature correlation rules and patterns in the data domain. The derived knowledge graph and data/models behind

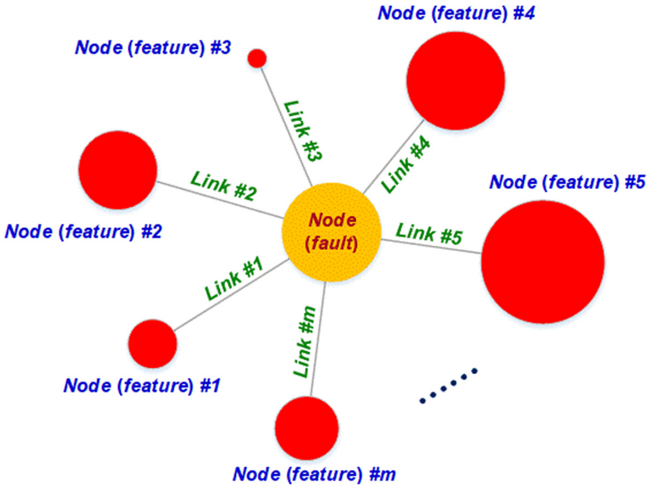

Fig. 2. Local view of a knowledge graph with respect to a fault mode.

it consolidate the knowledge base of the operators and maintenance staff to improve the field operation and management.

4 Results and Analysis

4.1 Datasets and Experimental Platform

The proposed system framework as Fig. 1 has been implemented using real case data sets from the high-speed railway administration. Since 2015, field data log files of on-board train control equipments used in practical railway operations were collected to build the big data-based system, where both the 200H and 300T on-board equipments were involved. In the laboratory, the big data platform was established using a master server and two slave servers. The four layers mentioned in Fig. 1 have been realized with specific software tools. Through the data pre-processing, the structured data sets, including all the features and fault mode labels, can be utilized in model training and derivation of the knowledge graphs. The procedures of knowledge graph construction based on the big data platform are described in Fig. 3.

The continuous accumulation of the filed data sets gradually consolidates the foundation of KG construction by this platform. By the end of 2019, there had been over 80 thousand log files of 200H on-board train control equipments, and the data amount almost reached 1TB. By integrating the Spark framework into the big data-based platform, the XGBoost algorithm can be carried out using certain training samples within a specific period of time. Though there have been approaches for XGBoost to determine a suggested number of trees, a fixed number 500 has been adopted for simplicity in generating the decision trees in the case study. Using the derived training model by XGBoost, the statistical results of different faults and the corresponding feature importance data enable us to construct or update the knowledge graph with respect to the faults need to be concerned in maintenance.

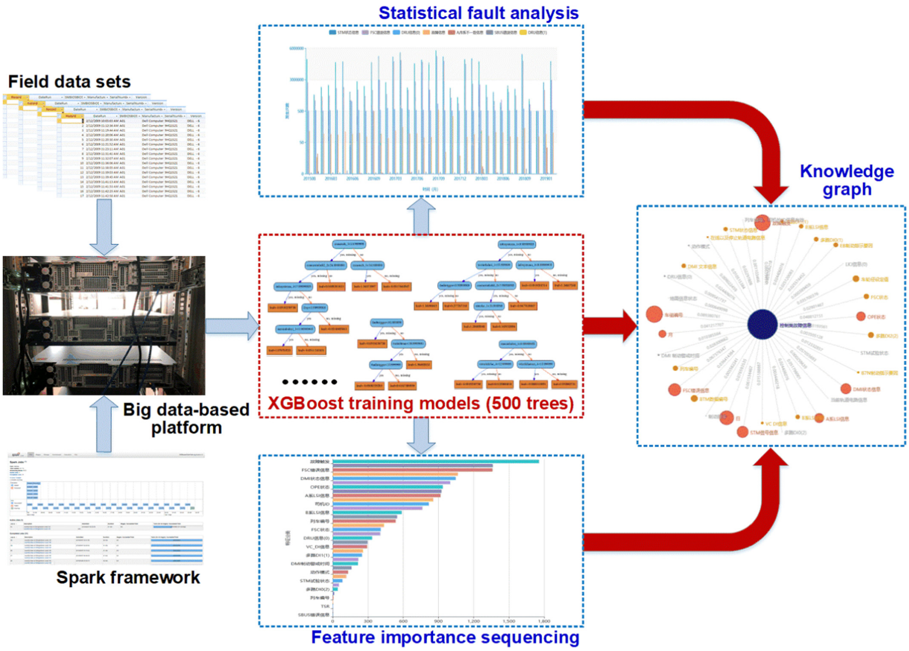
Statistical fault analysis
Field data sets
Knowledge graph
Big data-based platform
XGBoost training models (500 trees)
Spark framework
Feature importance sequencing

Fig. 3. Realization of big data platform and knowledge graph derivation.

4.2 Results of a Case Study

A case study is performed using the field data sets of 200H equipments from the January to June in 2017. The size of the adopted training set exceeds 131GB. It took a long time to carry out the XGBoost-based model training jobs. Figure 4 and Fig. 5 show the results of feature importance corresponding to the controlling fault mode.

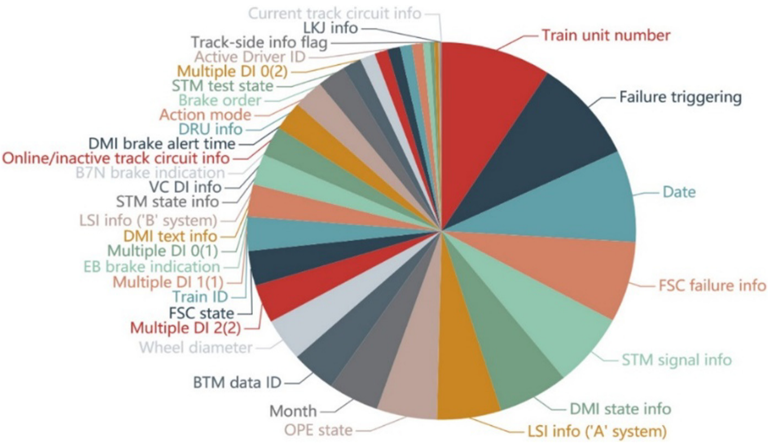

Fig. 4. Importance distribution of 33 major controlling-fault-related features.

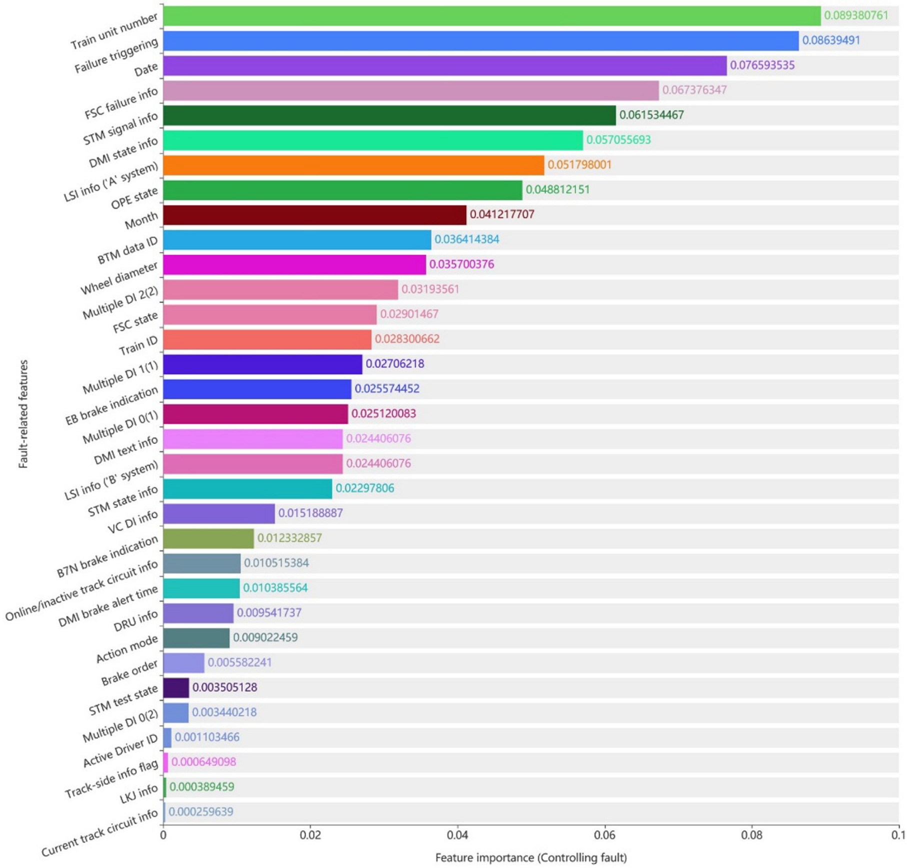

Fig. 5. Feature importance sequence of the controlling fault.

It can be found that not all the features are covered in the presented importance sequencing result. Based on the significance principle, only 33 major features from the whole feature set are considered in the feature importance sequencing and distribution analysis, since they achieved a contribution of 97.30% to the target fault mode. Figure 4 clearly shows the distribution of importance by each individual feature in the subset. In addition, Fig. 5 illustrates the differences among the 33 major features according to the normalized value of feature importance.

Based on the results of feature importance sequencing, the knowledge graph for the controlling fault mode is derived using the 33 major features. Using the open-sourced ECharts visualization tool, 33 feature nodes are integrated into this knowledge graph around a central node with a label "controlling fault". In order to distinguish the contributions to the fault mode from different features according to the machine learning model, the size of each feature node is determined dynamically according to the feature importance value. Furthermore, three colors have been adopted with a piecewise defined principle. That means a feature with a normalized importance larger than 0.05

indicates a red node in the derived KG, and the blue and green nodes represent the feature importance intervals of (0.01, 0.05] and (0, 0.01], respectively (Fig. 6).

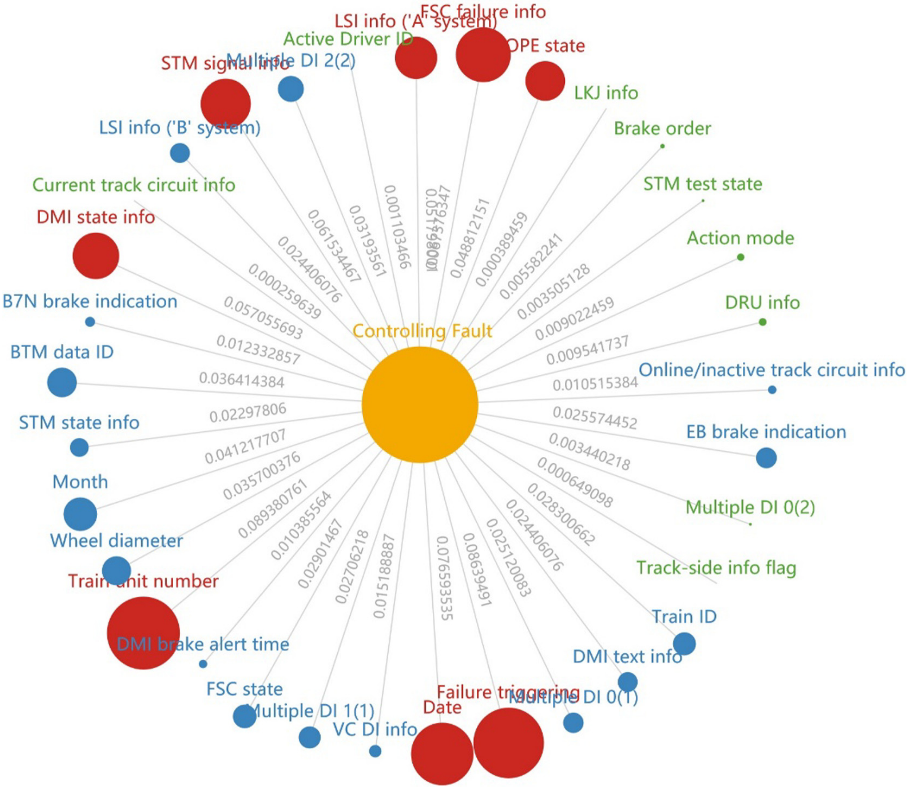

Fig. 6. Knowledge graph derived by the training model using the data set from 2017. (Color figure online)

We can find that there are 8 red nodes representing a higher importance. The other 16 blue nodes and 9 green nodes illustrate the moderate or lower influence to the fault mode by these features. The normalized importance values of each feature have been marked with the links. The knowledge graph clearly demonstrates a visualization interface to the users, which effectively reflects the inherence knowledge from a large data set. With the graph, we can use the central node “controlling fault” as the start to identify the features with significant effects to this fault mode according to the size, color and link value. The fault relevance information can be provided to specific engineers or staff to make decisions for adjusting maintenance plans. In addition, the prediction of fault probability can be achieved with new-coming feature data based on the trained model with respect to the knowledge graph, and that is a significant factor to enable the precise and effective condition-based maintenance. In general, the utilization of the XGBoost machine learning method based on a big data condition makes it possible of realizing the prognostics and health management of the specific on-board train control equipments based on the data-enabled knowledge of fault occurrence and progression over time during the life cycle.

It has to be noticed that only the "controlling fault" is involved in this case study. A complete knowledge graph corresponding to a whole on-board train control system level will cover all the defined fault modes. That means each feature will be connected with multiple sub-central nodes indicating the specific fault modes through several links. Thus, we will have to identify multiple attributes for a feature entity based on its contribution to different faults in constructing the knowledge graph and generating recommendations and assistance for the CBM purpose. Furthermore, only a fixed data set within half a year is adopted in the case study to demonstrate the knowledge graph constructing solution. Actually, this platform is capable of updating the derived models and knowledge graphs dynamically based on incremental data sets, which ensures the capability in life cycle monitoring and proactive maintenance of the on-board train control equipments in practical operations.

5 Conclusion and Future Works

This paper presents a data-driven framework for constructing the knowledge graph in pursuing the prognostics and health management of on-board train control system. A big data-based platform is designed and established to realize the storage and management of accumulative data files in the field operation. The XGBoost algorithm has been adopted to enable the model training with respect to specific fault modes, and it evaluates the importance of multiple feature entities. We also suggested principles of the knowledge graph with different nodes and links according to the derived feature importance. The design was realized with field data from real high-speed train control equipments. Results from a case study illustrate that the presented solution can determine the feature importance sequence and generate the knowledge graph with respect to specific fault modes, which reveals the relevance of the feature entities to the faults from a big-data-driven perspective.

In the future works, comparison analysis will be carried out based on multiple data sets within different time ranges, with which the effect of the incremental data condition will be further validated. Based on that, different training models will be used to perform fault probability prediction to realize a tight connection between models and decision making results for maintenance. Furthermore, more types of on-board train control equipment will be taken into account to enhance the coverage performance of the platform for assisting the practical operation and maintenance activities.

Acknowledgement. This research was funded by Beijing Natural Science Foundation (L191014), National Natural Science Foundation of China (61873023, U1934222), and Beijing Nova Program of Science and Technology (No. Z191100001119066).

The authors would like to thank Mei-hao Yang, Xin Han and Tian-tian Li for their assistance to the development and optimization of the big data-based platform and the cooperation with the operation maintenance department of the high-speed railway.

References

1. Saki, M., Abolhasan, M., Lipman, J.: A novel approach for big data classification and transportation in rail networks. IEEE Trans. Intell. Transp. Syst. **21**(3), 1239–1249 (2020)
2. Oneto, L., et al.: Train delay prediction systems: a big data analytics perspective. Big Data Res. **11**, 54–64 (2018)
3. Ghofrani, F., He, Q., Goverde, R., Liu, X.: Recent applications of big data analytics in railway transportation systems: a survey. Transp. Res. Part C: Emerg. Technol. **90**, 226–246 (2018)
4. Gulijk, C., Hughes, P., Figueres-Esteban, M.: Big data risk analysis for rail safety. In: Proceedings of 25th European Safety and Reliability Conference, Zurich, Netherlands, pp. 1–8 (2015)
5. Jiang, Z., Hsu, C., Zhang, D., Zou, X.: Evaluating rail transit timetable using big passengers' data. J. Comput. Syst. Sci. **82**, 144–155 (2016)
6. Fink, O., Wang, Q., Svensen, M., Dersin, P., Lee, W., Ducoffe, M.: Potential, challenges and future directions for deep learning in prognostics and health management applications. Eng. Appl. Artif. Intell. **92**, 1–15 (2020)
7. Chi, Z., Lin, J., Chen, R., Huang, S.: Data-driven approach to study the polygonization of high-speed railway train wheel-sets using field data of China's HSR train. Measurement **149**, 1–12 (2020)
8. Shahabinejad, M., Khabbazian, M., Ardakani, M.: An efficient binary locally repairable code for hadoop distributed file system. IEEE Commun. Lett. **18**(8), 1287–1290 (2014)
9. Bui, D., Hussain, S., Huh, E., Lee, S.: Adaptive replication management in HDFS based on supervised learning. IEEE Trans. Knowl. Data Eng. **28**(6), 1369–1382 (2016)
10. Yan, H., Yang, J., Wan, J.: KnowIME: a system to construct a knowledge graph for intelligent manufacturing equipment. IEEE Access **8**, 41805–41813 (2020)
11. Chen, T., Guestrin, C.: XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, USA, pp. 785–794 (2016)
12. Song, K., Yan, F., Ding, T., Gao, L., Lu, S.: A steel property optimization model based on the XGBoost algorithm and improved PSO. Comput. Mater. Sci. **174**, 1–13 (2020)
13. Chen, W., Fu, K., Zuo, J., Zheng, X., Ren, W.: Radar emitter classification for large data set based on weighted-XGBoost. IET Radar Sonar Navig. **11**(8), 1203–1207 (2017)
14. Georganos, S., Grippa, T., Vanhuysse, S., Lennert, M., Shimoni, M., Wolff, E.: Very high resolution object-based land use-land cover urban classification using extreme gradient boosting. IEEE Geosci. Remote Sens. Lett. **15**(4), 607–611 (2018)
15. Friedman, J., Hastie, T., Tibshirani, R.: Additive logistic regression: a statistical view of boosting. Ann. Stat. **28**(2), 337–407 (2000)