# Design of Intelligent Integration System for Multi-source Industrial Field Data Based on Machine Learning

Shufeng Zhuo[1(✉)] and Yingjian Kang[2]

[1] The Internet of Things and Artificial Intelligence College, Fujian Polytechnic of Information Technology, Fuzhou 350003, China
87742771@fjpit.edu.cn
[2] Beijing Polytechnic, Beijing 100016, China

**Abstract.** Aiming at the problem that information is scattered and difficult to manage in the current integration system, a multi-source industrial field data intelligent integration system based on machine learning is designed. The multi-source industrial field data synchronization device is designed, and the middleware technology is used to realize the integration of the field database, so as to realize the transparent access of the user to the field data source. Using machine learning-based host technology to integrate on-site data, design an intelligent retrieval engine for on-site data, and provide an integrated environment for users' data processing. Design data integration channel point-to-point circuits, independently select power lines, remove impulse noise, and facilitate visual data integration. Use machine learning methods to train weight parameters and build an integrated task scheduling model to minimize construction queuing to process extraction and operation and maintenance tasks. Adjust the data topology structure, according to the specific needs of multi-source industrial field data intelligent integration, use database connection pool technology to integrate field data, and check the integrity of the integrated data. It can be seen from the experimental results that the system integration effect is good.

**Keywords:** Machine Learning · Multi Source Industry · Site Data · Intelligent Integration

## 1 Introduction

In recent years, with the rapid development of automatic control technology and network technology, more and more on-site monitoring systems and data acquisition systems have been used in various links of enterprise production processes. These systems have improved the automation of the production process, and also provided more and more abundant production and economic data from the site for enterprise management [1]. However, today's on-site monitoring system and data acquisition system divide the real-time data of each link of the site into islands of real-time information, which makes it


B. Wang et al. (Eds.): ICMTEL 2023, LNICST 534, pp. 92–109, 2024.
https://doi.org/10.1007/978-3-031-50577-5_8

extremely difficult for enterprises to centralize and utilize real-time information [2]. The real-time data of each link of the enterprise is a part of the whole production process. If it is not centralized, it cannot effectively reflect the overall situation of production. The traditional data integration system is to read the meter at each production link regularly by the instrument workers, and then gather them. The workload is large, the situation is not reflected in time, the data synchronization is poor, and the informatization level of the enterprise production site is low, which greatly affects the improvement of the enterprise management level. The intelligent integration systems that can be applied now are mainly the integration system based on TRS network data radar and the integration system based on web spider software. These systems are basically similar in architecture, but the data search accuracy is low and the price is expensive. Based on this, an intelligent integration system of multi-source industrial field data based on machine learning is designed.

## 2 The Overall Structure Design of the System

In practical work, it is often encountered that the load suddenly increases, the load current increases, the load flow of the original cable is insufficient, and the over-current operation increases sharply. In order to increase the capacity, the cost can be saved by fully considering the operation of multiple cables in the original multi-source industrial site and laying cables again [3]. However, this method greatly increases the field data, so it is necessary to integrate the data. Based on the field operation system, an intelligent integration method of multi-source industrial field data based on machine learning is proposed to realize the coupling relationship between field systems under the integration system and ensure the efficient operation of multi-source industrial field. The overall structure of the system is shown in Fig. 1.

It can be seen from Fig. 1 that the multi-source industrial field data intelligent integration framework is composed of five layers, which are software support layer, integrated bus layer, data bus layer, service layer and application layer. The software support layer includes the underlying communication components for multi-source industrial on-site operation; the integrated bus layer provides a standardized interaction mechanism for the multi-source industrial internal modules, using a public object request proxy architecture [4]; The data bus layer is responsible for connecting the data platform and database access interface and data interface; the service layer can meet the platform interface of each application requirement, realize the relative independence of the service layer and the market business, and will not be affected by the change of the business process; The application layer can provide the application service interface required for the operation of the market [5]. The realization of specific functions between different layers can provide application services for the upper layer, and each layer has a clear fixed calling interface. The multi-source industrial field operation framework, as an industrial market data information release platform, integrates relevant real-time data to determine the coupling relationship and acts as a data sharing link.

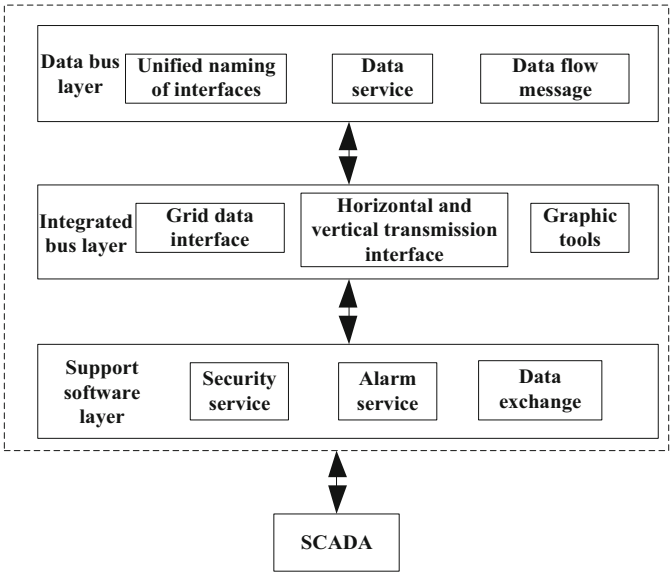



**Fig. 1.** The overall structure of the system

## 3 System Hardware Structure Design

The research goal of the intelligent integration system of multi-source industrial field data based on machine learning is to construct a transparent global data view required by users to support the global application of various databases and flexible data sharing among field databases on the basis of minimizing the impact on local autonomy of distributed field databases.

### 3.1 Multi-Source Industrial Field Data Synchronization Device

Use middleware technology to realize on-site database integration. The middleware is located between the on-site database system (data layer) and the application program (application layer). It coordinates each database system downward and provides a unified data schema and data upward for applications accessing integrated data. Accessing the common interface, the applications of individual databases still complete their tasks [6]. The data integration system is mainly composed of 4 parts including concentrator, adapter, metadata and management configuration, and WEB server. The system supports virtual views or view sets, unifies the data communication standards of the on-site database system, and solves the problems of data access, data extraction, data conversion, data integration and data display by using the background method, and realizes the user's transparent access to on-site data sources [7].

After the user submits a query, the concentrator translates the user query into one or more queries on the database. Then these sub queries are sent by the relevant adapter to the background database (field database) for query operation. The concentrator then processes the query results of each data source sent back by the adapter comprehensively.

Through the design of the integration mode, the relevant multiple data are integrated into one record, which is output and returned to the user through the same interface [8]. From the perspective of client and server, the middleware encapsulates the business logic of the system, and is built between the database service system and the application, forming a three-tier client server structure. Each field database resource constitutes the system data layer; The middleware system provides business services for data integration, constituting the business logic layer of the system; Applications constitute the presentation layer or transaction application layer of the system [9].

Realizing on-site database integration is to maintain the relative independence and autonomy of each application database system. After the unified comprehensive database is established, after the data in the distributed application database changes, the corresponding data in the comprehensive database should also change accordingly., only the incremental data of the application system that has changed is sent to the comprehensive database for synchronization.

### 3.2 Field Data Intelligent Retrieval Engine

The retrieval engine adopts the host technology based on machine learning to integrate the field data. When the memory capacity of the machine learning host exceeds the rated capacity, the driver of the application program will interrupt the data transmission with the memory host, thus providing a more stable data integration environment for users' data processing. The search engine structure is shown in Fig. 2.

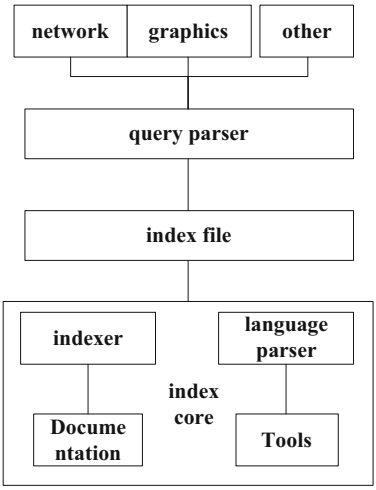



**Fig. 2.** Field data intelligent retrieval engine

It can be seen from Fig. 2 that the structure includes four parts: data search, data index construction, data storage and user interface. Data search is mainly responsible for searching the data most relevant to the field data from the database; Data index construction is mainly responsible for extracting index items from the searched data and

building index tables; Data storage is responsible for quickly finding relevant documents on the network and evaluating them; The user interface is responsible for querying user information [10].

### 3.3 Integrated Channel Circuit Design

The multi-source industrial field data integration channel point-to-point circuit takes a crystal oscillator as the core, and uses a balun circuit composed of capacitors and inductors, plus a whip antenna, which can transmit and receive wireless data. The node circuit diagram is shown in Fig. 3.

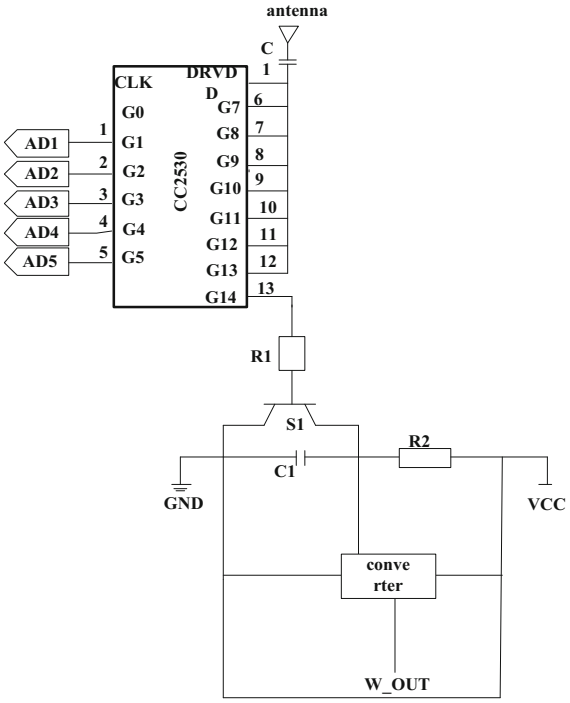



**Fig. 3.** Schematic diagram of the point-to-point circuit of the data integration channel

It can be seen from Fig. 3 that the control chip and LCD module components use 3.3 V DC voltage power supply, the power module uses three No. 5 batteries as the power supply, and the low-voltage differential regulator MCP1700–3.3 voltage is used to provide stable 3.3 V power supply. CC2530 single chip microcomputer is used as the main control chip, with an enhanced 51 core and 256 kB flash memory. The external equipment resources are very rich. Two crystal oscillators are externally connected, of which 32MHz crystal oscillator can provide accurate symbol period for wireless receiver. The button that controls the switch of the circuit is a non-self-locking independent button. In order to eliminate jitter, the program scans the button every 10ms. After storing the

key data, press the reset button to clear the variable and start a new round of transmission [11–12]. The ST8500 programmable power line communication system-on-chip is used as the core system of the modulation and demodulation module to independently select the power line and provide a reliable path for information reception. Design coupling ports with internal impedance to remove impulse noise for more accurate data integration results. Design point-to-point circuits to facilitate visual data integration.

### 3.4 Real-Time Database Construction

The function of real-time data system is the knowledge base of data, which provides efficient data storage, completes information query and operates real-time transactions, which is the basic unit of real-time database. Unlike ordinary commercial databases, real-time databases have the following characteristics:

(1) Time constraints

The main feature of real-time database systems is that time constraints are imposed on data objects and transactions. The time constraints and constraints on data are not only the ordinary consistency requirements of the database, but also temporal consistency requirements.

(2) Transaction scheduling

In traditional database systems, the goal of transaction scheduling is to improve the system's transaction throughput, but real-time database systems require that as many transactions as possible be completed within their deadlines. Therefore, the scheduling of real-time transactions is different from that in traditional database systems. Most of the real-time transaction scheduling strategies focus on the priority of transactions [13–14].

(3) Real-time data storage management of real-time database

The real-time database is mainly responsible for the storage and management of all real-time data in the system, and provides fast and accurate real-time information for related functions. Therefore, for the real-time database, real-time is the first priority. Considering this, the real-time database is in the system. During operation, it should occupy a small space and be resident in memory to ensure fast database reading, flexible access, and easy data sharing between functional modules.

When designing the real-time database, the following functional modules should be considered: the real-time database initialization module: This module is mainly used to create each data object based on the data required by advanced control, use linked list as the storage method, and establish the object name index corresponding to each data object, so as to improve the access speed of the data object and establish the historical database [15].

Basic operation module: Provides basic operations of data objects, such as obtaining other attributes of the data object through the name or ID of the data object, or obtaining the ID of the data object through the name and so on.

Read and write data operation module: Provide read and write data operations of data objects, write the field value stored in the data buffer into the field value attribute of the data object in the real-time database, and read the current value in the data object.

Communication device read/write operation module: manage the communication device, read the current working state of the device, and operate the specified device.

Window operation module: Read the name of the user window, operate the specified user window, and read the current state of the user window.

Alarm operation module: store alarm information and read the alarm limit of data objects.

Save operation module: store the data that needs to be saved in the SQL SERVER database.

## 4 Research on Key Technologies of the System

The machine learning network is based on the Bayesian probability generation model, which trains and adjusts the weight parameters between the hidden layer and the visible layer, so that the entire network can generate target data with maximum probability. Multilayer Boltzmann sets constitute a machine learning network. The neural network divides neurons into dominant neurons and recessive neurons. There are associative memory units between the upper and lower neurons. This connection has no direction and is used to realize associative memory function.

### 4.1 Integrated Task Scheduling Based on Machine Learning

The training and learning of a machine learning network consists of two steps: unsupervised training and tuning. The specific task of the unsupervised training step is to train a restricted Boltzmann machine in layers, with the output of each layer serving as the input to the upper layer. When the upper neurons are marked, joint training of marking is required; the tuning is mainly carried out in two stages: the cognitive stage and the production stage. In the recognition stage, the machine learning network generates the output of each layer layer by layer according to the input feature information, and uses the gradient descent method to correct the weight parameters of each layer. The basic state information consists of the top-level label annotations in the generation stage and the downward weight information, and the upward weight information is also modified in this stage.

In the process of feature extraction of machine learning networks, the input signals need to be represented by vectors and trained. The highest level associative memory unit divides tasks according to the clues provided by the lower level. Using the feedforward neural network based on labeled data, the machine learning network can accurately adjust the classification performance and recognize in the last layer of training. Compared with the direct use of feedforward neural network, this method has higher efficiency, because the machine learning network can be trained locally only by modifying the weight parameters, so the training speed is fast and the convergence time is short. Based on this, an integrated task scheduling model is built, as shown in Fig. 4.

It can be seen from Fig. 4 that an auxiliary variable is introduced to produce a loss queue. At the initial moment, this loss queue represents the difference between the current actual income of the system and the target. Introduce adjustable parameters to control the balance of system revenue and system delay to minimize construction queue processing

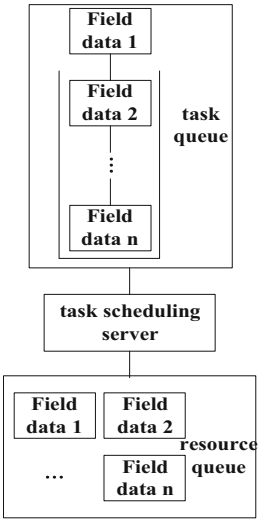



**Fig. 4.** Integrated task scheduling model

extraction and operation and maintenance tasks. Use the resource clustering algorithm to cluster resources, and generate scheduling task queues according to the resource clustering results, so that the allocation of resource queues is accurate and reasonable; for a queue, select the appropriate auxiliary resources, and select the maximum value; generate resources according to the results obtained in the calculation process Allocation vector; Update task vector, resource vector and resource allocation vector respectively according to the generation relationship between different formulas.

### 4.2 Data Topology Adjustment

According to the above integration task scheduling results, a multi-source industrial field data mapping table is formed, and intelligent data integration is completed through offline and online data adjustment.

#### 4.2.1 Multi-Source Industrial Field Data Mapping Table

The mapping table is the most important connection bridge between offline and online data, which limits the mapping of a large number of multi-source industrial field data to a large extent, and facilitates the maintenance of multi-source industrial field data. Based on the above analysis, the multi-source industrial field data mapping table mainly includes: generating a data mapping table in combination with timing characteristics; forming a field device mapping table; forming an AC line mapping table; forming a line mapping table. According to the above mapping table, the offline and online data are verified.

### 4.2.2 Offline and Online Data Verification

Use the node/branch power flow formula data verification tool to filter online damaged data. The process is as follows:

Assuming that $p$ and $o$ belong to the equivalence relationship in the multi-source industrial site database, the $p$ positive fields of $o$ are marked as $pos_p(o)$, which is expressed as:

$$pos_p(o) = \bigcup_o pXo \tag{1}$$

Based on formula (1), divide the multi-source industrial site data into a set in $o$, and calculate the important value $T(Y)$ of the $i$-th conditional attribute. The formula is:

$$T(Y) = \frac{pos_{X-1}(Y)}{pos_X(Y)} \tag{2}$$

where, $pos_X(Y)$ represents the attribute importance of category $X$ multi-source industrial site data, and $pos_{X-1}(Y)$ represents the attribute importance of category 4 multi-source industrial site data.

The larger the value of $T(Y)$, the greater the attribute weight of multi-source industrial site data. Then the weight value of multi-source industrial site data can be calculated by using the weight formula:

$$O_i = \frac{(t_X(Y) - t_{X-i}(Y))}{\sum_{i=1} \left[(t_X(Y) - t_{X-1}(Y))\right]} \tag{3}$$

Among them, $t_X(Y)$ represents the attribute weight of category $X$ multi-source industrial site data, $t_{X-1}(Y)$ represents the attribute weight of category $X-1$ multi-source industrial site data, and $t_{X-i}(Y)$ represents the attribute weight of category $i$ multi-source industrial site data.

The clustering distance of multi-source industrial field data can be calculated according to formula (4):

$$d'_{u'v'} = \sum_{t=1} \sqrt{O_u (y_{u'} - y_{v'})^2} \tag{4}$$

where, $y_{u'}$ represents the dimension value of multi-source industrial site data $u'$, and $y_{v'}$ represents the dimension value of multi-source industrial site data $v'$.

The data after clustering is processed, and the time series characteristics are used for comprehensive analysis, so as to build the abnormal judgment matrix of multi-source industrial field data:

$$Z(x) = \begin{bmatrix} z_{11} & \cdots & z_{1n} \\ \vdots & \ddots & \vdots \\ z_{m1} & \cdots & z_{mn} \end{bmatrix} \tag{5}$$

If the above matrix is met, the multi-source industrial site data is abnormal data, and the obtained multi-source industrial site data needs to be processed. The formula for filtering multi-source industrial site data is:

$$Z(t') = \frac{Z(x)}{\left(t_0' - 3\varepsilon, t_0' + 3\varepsilon\right)} \tag{6}$$

Among them, $Z(t')$ represents the normal data set obtained after multi-source industrial field data verification, $\varepsilon$ represents the filter, and $t_0'$ represents the local global time.

From the perspective of the whole network, the voltage phase angle measurement is less. According to the timing characteristics, the connection nodes with large parameter errors are estimated, and the parameters with small errors are input into the multi-source industrial field database, which can obtain accurate data and improve the quality of offline and online data.

### 4.2.3 Data Topology Adjustment

The offline data topology can only be changed manually, while the online data topology needs to be changed. By establishing a small branch to combine multiple computing nodes, and setting the components at both ends to reflect the disconnection of the equipment, the online data topology can be changed.

## 4.3 Integration Process Design

According to the specific requirements of intelligent integration of multi-source industrial field data, relevant information is collected and processed. Based on the characteristics of time series, the data integration library is constructed with the maximum membership, and the gravity center method is used for comprehensive analysis. The time series characteristics are used for comprehensive analysis to build the analysis matrix:

$$X = \begin{bmatrix} x_{11} \ \ldots \ x_{1n} \\ \vdots \ x_{ij} \ \vdots \\ x_{n1} \ \cdots \ x_{nn} \end{bmatrix} \tag{7}$$

The quantitative formula (7) converts the quantitative result into an integer value between 0 and 1 for users to query, and data retrieval is the premise of integration. In-depth analysis of the advanced retrieval function of the database will increase the storage space of the database.

Use the binary data conversion method to match the extracted data. The specific matching process is as follows:

(1) Determine the convex or concave growth relationship between adjacent points through the binary sequence between adjacent points, and use the analytic hierarchy process to calculate and formulate the subjective set weight vector $\omega$:

$$\omega = [\omega_1, \omega_2, \cdots, \omega_n] \tag{8}$$

According to the characteristics of offline and online data flow, the calculation method of anti-entropy weight method is used to obtain the difference matrix:

$$G=\begin{bmatrix} x_{11}\omega_1 \ \ldots \ x_{1n}\omega_{1n} \\ \vdots \ \vdots \\ x_{n1}\omega_{n1} \ \cdots \ x_{nn}\omega_{nn} \end{bmatrix} \tag{9}$$

According to the analysis matrix, calculate the difference value between different data, set the weight vector, and reasonably design the data dictionary.

(2) The trend proportional data reduction method is used to reduce the candidate sequence and pattern between adjacent points of the determined relationship to the same interval. Data reduction is a basic operation to convert attribute sets into target strings through insert, delete and replace operations. When two attribute sets are converted to the target string form, the distance formula between the converted two strings expressed by the edit distance is:

$$d_{x_i y_j}(i,j) = \min \begin{cases} d_{y_j}(i,j-1)+1 \\ d_{x_i}(i-1,j)+1 \\ d_{x_i y_j}(i-1,j-1)+1 \end{cases} \tag{10}$$

In formula (10), $x_i$ and $y_j$ represent the string conversion results of attribute sets $c_i$ and $c_j$, respectively; $d_{y_j}(i,j-1)$ represents the string $c_j$ to delete a number; $d_{x_i}(i-1,j)$ represents the string $c_i$ to delete a number.

The reduced data is regarded as independent distributed sample data of machine learning, and the attribute set relationship between any two data is analyzed.

(3) Calculate the sequence similarity in the same interval, so as to distinguish the amplitudes of convex growth or concave growth with different change amplitudes. According to the amplitude, obtain the set of sub sequences, that is, the pattern matching result.

According to the machine training results, the attribute sets of $x_i$ and $y_j$ are $c_i$ and $c_j$. Let $\alpha$ be the path of $c_i$, $\beta$ is the path of $c_j$, $\overline{h_1}$ is a vector of data $x_i$ derived from $\alpha$, and $\overline{h_2}$ is a vector of data $y_j$ derived from $\beta$. The similarity of two vectors can be expressed as:

$$Simx\left(\overline{h_1}, \overline{h_2}\right)=\omega' \times \frac{2 \times x(G(\alpha,\beta))}{x(\alpha)+x(\beta)} \tag{11}$$

In formula (11), $x(G(\alpha,\beta))$ represents the longest common sub path length of path $\alpha$ and $\beta$; $\omega'$ is the weight coefficient of similarity; $i,j$ represents the $i$ and $j$ vectors respectively.

Based on this, calculate the similarity between two documents:

$$Sim\left(x_i, y_j\right)=\frac{\sum_{i=1}^{b_1}\sum_{j=1}^{b_2} Simx\left(\overline{h_1}, \overline{h_2}\right)}{b_1+b_2} \tag{12}$$

In formula (12), $b_1$ and $b_2$ represent the numbers of $x_i$ and $y_j$ data, respectively. The vectors fuse the underlying speech structure and content features, while also reflecting similar parts of the data. According to the calculation result, the sequence similarity in the same interval can be reflected.

According to the similarity calculation results, the database connection pool technology is used to integrate the field data. The connection pool is the storage pool of connection objects. The amount of information is controlled through the internal connection mechanism, and the query interface in the system structure is used to provide the connection channel. The API function is used to connect with the on-site database, which can effectively convert the query statements to the specific database, obtain various data of the on-site database, and generate mapping files. The file cycle for each mapping is:

$$t{=}t1 + t2 + t3 + t4 + t5 \tag{13}$$

In formula (7): $t1$ represents the mapping end time; $t2$ represents the network delay; $t3$ represents the timing logic setup time; $t4$ represents the skew of the mapping signal; $t5$ represents the network delay. Thereby, the minimum cycle for generating the mapping file is obtained, which improves the file mapping efficiency. To fully combine the personalized service mode, it is convenient for users to use it. It is necessary to organically combine relevant majors and put them in different folders to form integrated files.

### 4.4 Integrated Data Integrity Detection Based on Machine Learning

The integrated data is used as machine learning input data, and the integration behavior is defined by association rules. Through traversing each rule on the basis of rules, the data package load is detected.

Assume that there are multiple collections of distinct items, given a transaction database, where each transaction T is a collection of a set of items, that is, with a unique identifier. If the itemset is $X \subset T$, then the transaction set T contains the itemset X, and the implication form of $X \Rightarrow Y$ is the general representation of association rules. Association rules contain two parts, support and confidence. Among them, if a transaction also supports a group of transactions, it is called the support of association rules for a given total transaction database, that is, the support degree of association rule $X \Rightarrow Y$. Support supports the frequency of expression rules, and describes the antecedents and the proportion of antecedents in the entire dataset. Minsup is used to represent the minimum support, and the minimum support represents the minimum statistical significance of the data item set. An item set whose support is greater than the minimum support is called a frequent item set.

For the known total transaction data set, among the transactions supporting transaction set X, there are also transactions supporting transaction set Y, that is, the confidence level of association rule $X \Rightarrow Y$. Confidence represents the strength of the rule and describes the likelihood that the rule will occur if its preconditions are met. The minimum confidence level is represented by minconf to represent the minimum confidence level of the rule, and the association rule whose confidence level is greater than the minimum confidence level is called a strong rule. In the transaction database, the problem of association rule mining is to specify the value of minsup with the minimum support and

the value of minconf with the minimum confidence, and find the rules whose support and trust are both higher than the two predetermined values.

Based on the above rules, once found, an alarm message will be reported. In the detection process, the detection process is divided into two parts: multi-mode matching and verification. After the detection rules are loaded into the rule set, the rule set will be pre-combined, and the rules with the same characteristics are put into the same signature group, and the integrated data integrity detection is realized through rule matching.

Firstly, enter the multi pattern matching stage, input the data into the rule matching engine, and then recognize according to the data characteristics obtained. The message will first enter the first stage of detection, that is, the multi pattern matching stage. In the multi-mode matching stage, a rule matching the message characteristics is usually selected first to obtain the generated message to be detected and obtain the pre verification data set.

Then enter the verification phase. In the verification phase, the message is traversed one by one. When the message characteristics meet all the constraints of the rules, alarm information is generated. The detailed detection process is as follows:

Step 1: According to the time complexity of multi-pattern matching, candidate rules are extracted from multiple feature sets;

Step 2: For the multiple multi-pattern matching rules obtained in step 1, it is assumed that the number of multi-pattern matching rules is;

Step 3: Traverse the rules one by one. If the signature is set to “hit”, the n-1 times before the “hit” are invalid, so the position of the last signature in the signature sequence is determined, which is valid. Hit the result, the result is the integrated behavior data.

As part of the integrated data is related to multi-source industrial site construction, a three-level detection mode is required to ensure that no data will be missed during the integrated detection. The first level of detection is mainly used to detect the operation status of all multi-source industrial sites and search for abnormal operation status information; The second level of detection is mainly to respond to abnormal data. After the first level of detection, integrate all data, build an abnormal database, and determine different attack means through integration and classification; The third level of detection is to compare and analyze the first two levels of detection results with the relevant abnormal operation data in the historical database, and obtain all the analysis results. After that, machine learning and big data technology are used to query whether the behavior status is kept within the normal indicators, and then the integrity of abnormal operation behavior data is analyzed quantitatively and regularly through preprocessing to obtain accurate analysis results.

## 5 System Test

### 5.1 Experimental Platform

In order to test the performance of the multi-source industrial field data intelligent integration system based on machine learning, the data integration integrity of the system is compared with that of the integration system based on TRS network data radar and the integration system based on web spider. The experimental environment of the three systems is consistent, that is, the Windows host is i7-12700F; The processor is 13700K; The

data register is 74HC595D; The database host is Redis cloud database; The processing chip is LQFP100; The microcontroller is stc89c52RC; The timing regulator is PZ-51 Tracker; The simulation software is Matlab R2019a. In order to ensure the authenticity and reliability of the experimental results, it is necessary to obtain the experimental results by increasing the number of experiments and cross-validation. The integrated system platform is a platform used to monitor violations of system security principles in multi-source industrial sites, and its structure is shown in Fig. 5.

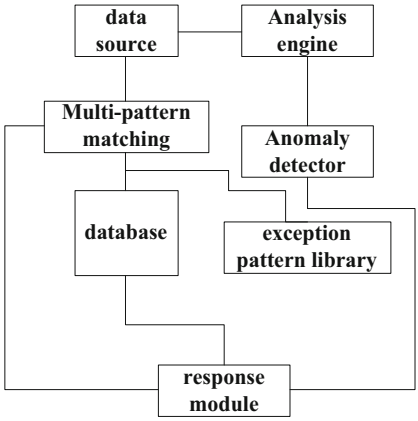



**Fig. 5.** Experimental platform

It can be seen from Fig. 5 that the data source provides monitoring data for the platform, and the analysis engine is responsible for analyzing and reviewing the data. Once abnormal integration results are found, the data will be immediately transferred to the response module. The response module responds randomly and generates feedback information.

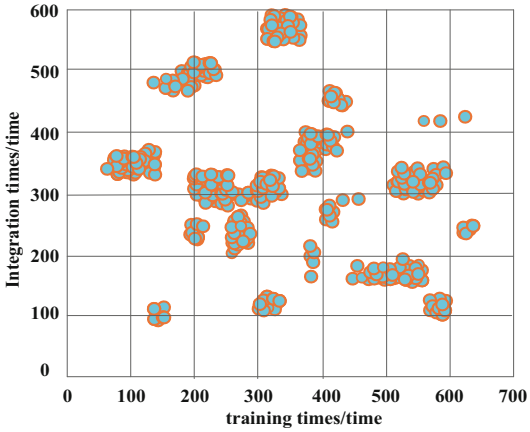
### 5.2 Experimental Indicators

Using the accuracy rate and recall rate of the data integration results as the evaluation criteria, the formula can be expressed as:

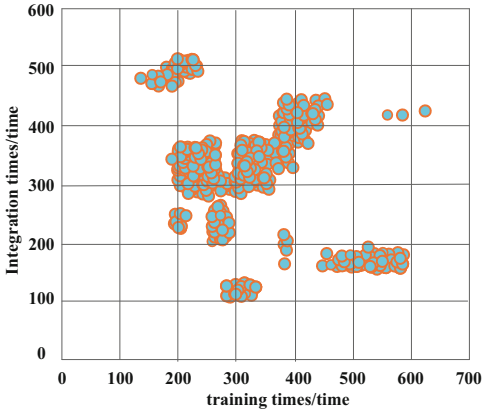
$$P=\frac{Q_A}{(Q_A+Q_B)} \tag{14}$$

$$R=\frac{Q_A}{(Q_A+Q_C)} \tag{15}$$

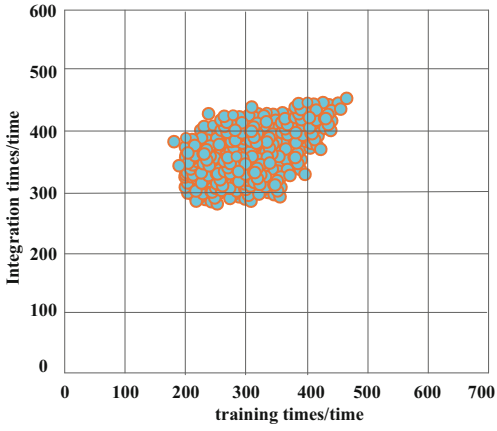
In formulas (14) and (15), $Q_A$ represents the same kind of documents whose semantic similarity calculation result is greater than zero and judged to be the same attribute; $Q_B$ refers to similar documents with semantic similarity calculation results greater than zero and judged as different attributes; $Q_C$ refers to documents of the same category whose semantic similarity calculation result is equal to zero and judged to be the same attribute. The larger the calculation results of formula (14) and (15), the more accurate the inspection results are.




(a) Integrated system based on TRS network data radarvv




(b) Integrated system based on web spiders




(c) Intelligent integration system based on machine learning

**Fig. 6.** Comparative analysis of data integration effects of different systems (a) Integrated system based on TRS network data radarvv (b) Integrated system based on web spiders (c) Intelligent integration system based on machine learning

### 5.3 Analysis of Data Integration Effect

For the analysis of data integration effect, the integration system based on TRS network data radar, the integration system based on web spider and the multi-source industrial field data intelligent integration system based on machine learning are used for comparative analysis. The comparison results are shown in Fig. 6.

It can be seen from Fig. 6 that with the integration system based on TRS network data radar and the integration system based on web spiders, the data is still in a decentralized structure, while with the integration system based on machine learning, the data integration effect is good.

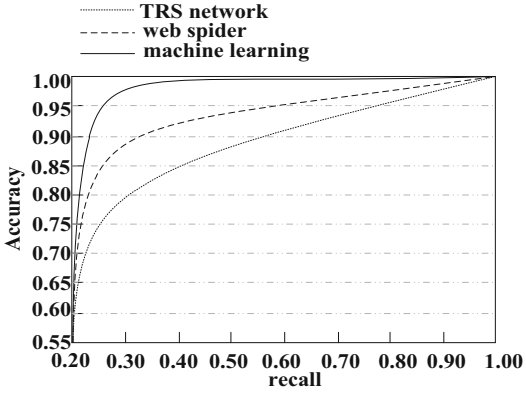
### 5.4 Data Integration Integrity Check

Firstly, set three external interference conditions, namely, buffer overflow, running process infection, and Trojan Horse. The root cause of buffer overflow is that C + + is insecure, and there is no limit to the reference between the array and the pointer. Buffer overflow vulnerabilities are very common, and exist in C language development tools; Running a process means that one process can read and write to another process. The attack mode is to directly insert the sbrk() function into the code, or find an appropriate space in the process space, then write code and rewrite the original data, so as to achieve the purpose of malicious code execution; A Trojan horse is a program that exists in a computer and can steal passwords, copy or remove files. Then select the experimental set as $r_1$ and $r_2$, and finally train the two experimental sets to get the ROC curve, as shown in Fig. 7.

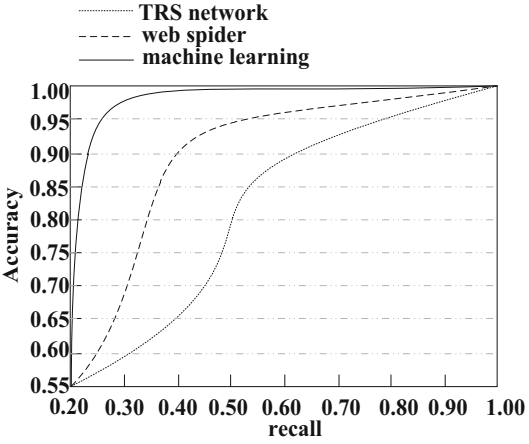
It can be seen from Fig. 7 that under the two interference conditions of the integration system using machine learning, the accuracy of the data integration results is always greater than 0.95 when the recall rate is greater than 0.40, and the accuracy of the other two methods is lower than that of the machine learning method. This shows that the use of machine learning-based integration systems can obtain fully integrated data.

## 6 Conclusion

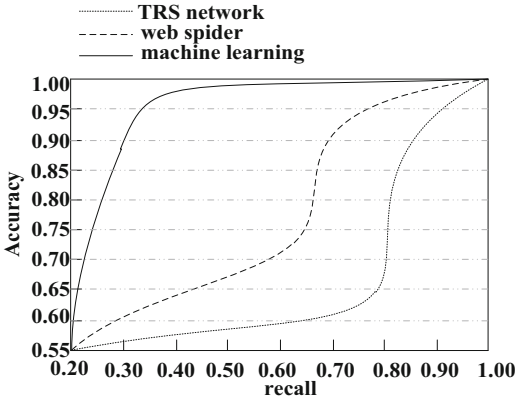
In view of the large part of noise in the data obtained by the current integration system, and the low data integrity, which leads to the serious problem of data loss in the integration process, the intelligent integration system of multi-source industrial field data based on machine learning is designed to solve the problems existing in the current system. Self-selected power lines are used to remove impulse noise, facilitate visual data integration, and combine multiple pattern matching rules to achieve integrated data integrity detection, so as to achieve high data integration. The system can effectively solve the problem of data loss in the integration process, but there is still a problem that needs further in-depth study, that is, for the frequent distributed combined network attacks, the essence of machine learning needs to be enhanced, so that the research method can adapt to the changing multi-source industrial field environment.




(a) Buffer overflow




(b) Running process infection




(c) Trojan Horses

**Fig. 7.** ROC curve (a) Buffer overflow (b) Running process infection (c) Trojan Horses

## References


1. Baoxue, L., Zoujing, Y., Chunhui, Z.: Variational imputation model of multi-source industrial data based on domain adaptation. Control Eng. China **29**(4), 627–636 (2022)
2. Ming, C., Jiyuan, Y.: Research on establishing Chinese altmetrics data integration and analysis platform. J. Acad. Libr. **40**(4), 110–119 (2022)
3. Sanning, H., Yuxiang, L.: Cross social network user matching method based on multi-source data integration. Comput. Simul. **38**(4), 352–355+466 (2021)
4. Long Beiping, W., Jiajie, Z.Q., et al.: A method of the real estate data integration based on weighted similarity model. Bull. Surv. Mapp. **6**, 122–126 (2021)
5. Sunfa, L., Zhixing, L.: Design of server-side data integration system based on virtualization technology. Mod. Electron. Tech. **43**(2), 77–79+83 (2020)
6. Xia, Y., et al.: Research on data integration of command information system based on SOA and ontology. Fire Control & Command Control **47**(3), 136–143 (2022)
7. Shuangqin, L., Rui, X., Wenchen, C., et al.: Design of time dimension big data flow integration system based on multi-dimensional hierarchical sampling. Mod. Electron. Tech. **43**(5), 133–136+140 (2020)
8. Wu, L., Li, Z., Abourizk, S.: Automating common data integration for improved data-driven decision-support system in industrial construction. J. Comput. Civil Eng. **36**(2), 4021037.1–4021037.17 (2022)
9. Liu, X.: Design of enterprise economic information management system based on big data integration algorithm. J. Math. **22**(10), 1–9 (2022)
10. Wang, K., Liu, X.: An anomaly detection method of industrial data based on stacking integration. J. Artif. Intell.Artif. Intell. **3**(1), 9–19 (2021)