Academic Journal of Applied Sciences
ISSN: 3105-0549 (Print), ISSN: 3105-0557 (Online) | Vol. 3, No. 3, 2026

Research on information network data cleaning and fusion technology based on Feature Engineering

Xin Zhao, Jianhong Yang, Ying Dong *

CATARC Information Technology (Tianjin) Co., Ltd., China

* **Corresponding author**: (Email: dongying@catarc.ac.cn)

Abstract: The on-board terminal, roadside unit, business system and other nodes are coordinated. The data presents the characteristics of multi-source heterogeneity, high noise and strong timing. The collection often introduces omissions, abnormalities, repetitions and inconsistencies, which affects the perception and decision-making of the upper level. This article takes the multi-source data cleaning and integration of the information network as the object, and proposes a data governance framework with feature engineering as the main line. The framework takes the four-dimensional quality evaluation of integrity, accuracy, consistency and timeliness as the starting point to locate defects; time alignment, missing interpolation, abnormal detection and repair and other operations are carried out in turn. Among them, the interpolation adopts the EM algorithm based on the expectation of multivariate Gausssian conditions, and the abnormal detection adopts the 3 σ criterion of de-trend residual. After cleaning, the special The variance inverse weighting and principal component analysis are used for grading. Experiments show that the feature engineering-oriented cleaning and fusion method can significantly improve the data quality of the information network and provide a consistent and usable data basis for subsequent analysis.

Keywords: Information network, Data cleaning, Feature engineering, Data fusion, Data quality.

1. Introduction

Relying on sensing, communication and computing facilities, the information network perceives, transmits and processes physical objects, which is widely used in intelligent transportation, industrial monitoring and urban governance. Its data comes from on-board terminals, roadside units, sensors and background systems, with sampling frequency, outline, accuracy and reliability. Affected by interference, failure and packet loss, the data is often missing, abnormal, duplicate and inconsistent. Direct use will distort state estimation and decision-making, leading to false reports, mischeduling and even accidents. Therefore, data cleaning and integration are the foundation of governance. However, the data flow must be completed within a limited time window, and the node computing power is very different, so it cannot be manually verified one by one. The method should use statistical structure, take into account accuracy and expenses, and maintain low complexity.

Data cleaning has been successful in many fields of engineering. For example, the literature [1] proposes an LSTM-based cleaning method for the abnormality and omission of high-speed control point retest data: construct a decision-making rule set with attribute correlation and importance to achieve simplication; use association rule mining and utility importance to find the timing correlation, and LSTM identifies abnormalities; according to the weight factor and the missing interval length Adapt to divide the pattern and fill in. Experiments show that it can reduce the loss rate and improve the data quality. Literature [2] In order to solve the problem that traditional statistical methods are difficult to efficiently detect and clean massive abnormal data, this paper builds an identification cleaning system that integrates unsupervised learning and attention mechanisms. Abnormal identification uses AutoEncoder to detect multiple types of anomalies, combined with 3D DBSCAN to extract long-term large-scale abnormal point clusters; Transformer is used for cleaning to learn rainfall and multi-source hydrometeorological mapping. Verification of the hourly rainfall data of 11,805 rainfall stations in the upper reaches of the Yangtze River from 2013 to 2024: 94% of the sites can effectively learn the time and space-time laws of rainfall, and identify 25 abnormal clusters lasting >30h and with a radius of >100km from January to July 2024. Transformer cleaning reached 0.96 in NSE in 2022 and 33.81dB in PSNR, but the estimate of heavy rainfall in 2023 needs to be improved. The framework has good potential in the range of light rain to normal moderate rain.

Feature engineering and data integration are two key links connecting cleaning results and downstream tasks. For example, the literature [3] is based on the design of an evaluation analysis model based on the deep neural network for the subjective experience of power engineering information data evaluation, such as multi-dependence hierarchical analysis. Build a MLSTM-GNN-BERT module to realize multi-source data fusion and feature extraction, with the ability to extract text, images and timing features; combine time convolution network and deep reinforcement learning to design a dynamic evaluation module to output project evaluation results. The atting and comparison experiments show that the accuracy rate of the model is 91.1% and the recall rate is 92.1%, which verifies the correctness of the design and the effectiveness of the module, which is better than the mainstream algorithm and has practical value in engineering. Literature [4] is based on the pain points of subjective experience such as over-reliance on hierarchical analysis of power engineering information data evaluation. This article builds an evaluation analysis model based on deep neural networks. The MLSTM-GNN-BERT module is used to complete multi-source data fusion and feature extraction, which can adapt to multi-type input of text, images and timing; integrate time convolution network and deep reinforcement learning to build a dynamic evaluation module and output project evaluation results. The ablation and comparative

experiments show that the accuracy rate of the model is 91.1% and the recall rate is 92.1%. The effectiveness of each module has been verified. The comprehensive performance is better than that of the mainstream algorithm, and it has engineering application value.

In summary, existing research often separates data cleaning, feature engineering, and data fusion: cleaning focuses on quality restoration, feature engineering focuses on representation construction, and fusion defaults to clean input data. The core difficulty of information network data governance lies in the strong coupling of the three, the constraint of cleaning effect on feature quality, and the determination of defect processing priority by feature and fusion strategy. Currently, there is a lack of integrated systematic research around the main line of feature engineering.

2. Data Feature Analysis

2.1. Multi-source And Heterogeneous Features

There are four characteristics of information network data: First, there are various sources. For example, the state of the same section in intelligent traffic is jointly described by the on-board terminal, roadside unit and background system, and the sampling cycle is different. The second is heterogeneity. The definition, scale and accuracy of the same physical quantity are different from each source, resulting in the measurement center not overlapping and the characteristic scale crossing the order of magnitude. Third, there is a strong timing. The observed value changes continuously and the adjacent is highly correlated. The event-driven data is suddenly accumulated, and the timing structure needs to be used. Fourth, the quality is uneven. The node equipment and the environment are very different. High reliability and low reliability sources coexist, and missing, abnormal, duplicate and inconsistency are often superimposed at the same time.

2.2. Data Quality Issues

Assuming there are K data sources in the information network and the observation time is $t = 1,2,\cdots,T$. The mathematical expression for data observation is shown in formula (1).

$$d_k^t = s_t + b_k + \varepsilon_k^t \tag{1}$$

In formula (1), d_k^t is the observation vector of the k-th source at time t, s_t is the true state, both satisfy the additive noise model, b_k is the fixed deviation of the k-th source, and ε_k^t is the random noise with a mean of zero and variance of σ_k^2.

Based on the mathematical expression (1), the four types of quality problems are defined as: first, missing values, that is, the observation component is not collected and manifests as empty values; second, abnormal values, that is, the observation is significantly distorted by failure or interference, and the deviation far exceeds the noise; third, the repeated value, that is, the same observation is recorded many times, the copy is the same or slightly disturbed; fourth, the data is inconsistent, that is, there is a system conflict between different sources for the same state description, including source-level deviation and time stamp misalignment.

2.3. Data Quality Evaluation Metrics

In order to quantitatively evaluate and compare the cleaning effect, this article defines quality indicators from the four dimensions of completeness, accuracy, consistency and timeliness. All values are normalized between 0 and 1. The closer to 1, the higher the quality. The proportion of non-empty records in the integrity measurement observation unit. If the data matrix D contains M observation units and the set of missing units is N, the integrity expression is defined as formula (2).

$$C = \frac{M - |N|}{M} \times 100\% \tag{2}$$

Accuracy measures the degree to which observed values deviate from the true state. Assuming the main observation after cleaning is $\overline{d_t}$ and the true state is d_t, the mean absolute error normalization formula is used as shown in formula (3).

$$\overline{MAE} = 1 - \frac{I}{M} \sum_{t=1}^{M} \frac{\left|\overline{d_t} - d_t\right|}{\frac{1}{M}\sum_{T=1}^{M} |d_t|} \tag{3}$$

Formula (3) defines the error as a relative scale based on the average absolute value of the true state, which facilitates cross scene comparison. Consistency measures the proportion of consistency described by different sources at a common observation point.

3. Data Cleaning Method Based on Feature Engineering

3.1. Overall Process

The data cleaning process is divided into: time alignment, missing interpolation, abnormal detection and repair, and duplicate record identification according to the order of data flow. Time alignment resamples each source to a unified time grid to eliminate timestamp drift; missing interpolation estimates fills empty values; abnormal detection, repair identification and replacement of distorted observations; duplicate identification deletes redundant copies. Each link uses multi-source redundancy, chronological distribution information, record key and field similarity respectively. The design follows two principles: first deal with defects that affect integrity, so time alignment and interpolation prefix; each link records the processing log, which is convenient for retroactive audit and ensures interpretability.

3.2. Time Alignment and Consistency

Time alignment maps each source observation to a unified time grid $t = 1,2,\cdots,T$. For observations that deviate from the grid at the sampling time, nearest neighbor or linear interpolation is used to estimate the value of the grid time. For the kth source, assuming its observation sequence is (t_k^j, d_k^j), the calculation formula for the value of grid time t after alignment is formula (4).

$$\overline{d_k^t} = d_k^{j^*}, \; j^* = \arg\min_j |t_k^j - t| \tag{4}$$

To avoid introducing excessive errors in time alignment, an

alignment tolerance is set: drift is retained within the tolerance, and any deviation is considered invalid and removed. Source level deviation calibration utilizes redundant information from multiple sources to estimate fixed deviations from each source, correct systematic offsets, and improve alignment accuracy and subsequent fusion reliability. This strategy balances fault tolerance and data quality. On a common observation point with no anomalies or missing data, the expected difference between the observations of data source i and data source j is equal to the difference in bias $b_i - b_j$. Therefore, the mathematical expression for bias estimation using the robust median of the difference is formula (5).

$$\overline{b_{i-j}} = median(d_i^t - d_j^t) \tag{5}$$

3.3. Missing Value Imputation Methods

Missing interpolation aims to estimate the missing value using observable information. The mean interpolation is simple but ignores the correlation and timing; the median interpolation is anti-outer but does not use the variable relationship; the regression interpolation uses the variable linear relationship, but it is difficult to handle when multiple columns are missing at the same time; the EM algorithm assumes the multivariate Gaussian distribution, iteratively estimates the mean and covariance, and expects conditional interpolation. For a comprehensive comparison, this article chooses EM as the main process method [5-7].

Firstly, assuming the complete data is $D = (X_{obs}, X_{mis})$ and the parameter set is $\theta = (\mu, \Sigma)$. The EM algorithm starts from the initial estimation and alternates between two steps.

Step 1: Step E calculates the conditional expectation of missing data under the current parameters, constructs the log likelihood expectation of complete data, and the expression of the log likelihood function is formula (6).

$$Q(\theta \mid \theta^{(r)}) = \mathrm{E}[\log L(\theta; D) \mid D_{obs}, \theta^{(r)}] \tag{6}$$

Step 2: Step M maximizes the expectation to update the parameters, iteratively until convergence, and the mathematical expression for updating the parameters is formula (7).

$$\theta^{(r+1)} = \arg\max_{\theta}(\theta \mid \theta^{(r)}) \tag{7}$$

Iterate until the parameter change is less than the threshold, and use the last E-step conditional expectation as the missing position interpolation value. EM interpolation can utilize the correlation structure of all variables and maintain a low error even when the missing rate increases, making it suitable for handling missing observations in multivariate information networks. Its effectiveness depends on two premises: the data is randomly missing, and the probability of missing is independent of the observed value; The data approximately follows a multivariate Gaussian distribution or approximately satisfies it after transformation. The loss of packets in information networks usually satisfies random loss, and the loss caused by extreme events needs to be identified through abnormal processing first. When the correlation between variables is weak, EM gain is limited and can be replaced by regression interpolation.

3.4. Anomaly Detection and Handling Methods

Abnormal detection requires distinguishing between normal noise and abnormal distortion. Time series signals in information networks are usually non-stationary, and the global mean and standard deviation are affected by trends. Directly using the global 3 σ criterion may result in missed detections due to a wide threshold, making it difficult to accurately identify anomalies. Therefore, detection strategies that can adapt to local changes need to be adopted [8-9]. This article adopts the detrended residual criterion: first estimate the local trend using the sliding window mean, and then apply the 3 σ criterion to the residual. The mathematical expression for the residual of window width w at time t is shown in formula (8).

$$r_t = d_t - \frac{1}{w}\sum_{j=t-w+1}^{t} d_j \tag{8}$$

In formula (8), if the residual satisfies $|r_t - \bar{r}| > 3\sigma_r$, it is considered abnormal [8-9], where r_t and σ_r are the mean and standard deviation of the residual, respectively. This criterion eliminates the tendency component to raise the threshold while retaining sensitivity to local mutations.

Local median replacement is used for anomaly repair: for the anomaly time t, the median of five normal observations before and after is taken as the repair value. Local median anti trend can maintain signal continuity. Detection and repair are performed on the complete data after interpolation to avoid missing interference. There is a compromise in fixing: the higher the recall, the more thorough the correction, but there are also more false positives and rewrites of normal points.

4. Data Fusion Methods

Data fusion is divided into data level, feature level and decision-making level according to the degree of abstraction. The data level directly combines the original observation, and the information loss is minimal, but it requires strict alignment and consistency; the feature level is combined in the feature vector layer, using unified expression to tolerate alignment errors and noise; the decision-making level, each source is independently judged and merged, which is the most flexible but may lose details. In view of the large differences in the source framework and accuracy of each information network, this paper adopts the feature-level integration method. Feature-level fusion is combined with the main characteristics of each source after cleaning. This paper examines four methods of equal weighted average, inverse weighting of variance, median and main component analysis.

Equal weight averaging directly takes the arithmetic mean of each source principal variable, which is simple to implement and requires less computation.

The reciprocal weighting of variance allocates weights based on the variance of each source noise. The smaller the noise, the greater the weight of the source, in order to suppress the influence of high noise sources and improve fusion accuracy. This method is computationally simple and suitable for scenarios where the variance of noise sources can be estimated and relatively stable. The weight calculation formula is shown in formula (9).

$$\bar{s} = \sum_{k=l}^{K} w_k \cdot d_k$$
$$w_k = \frac{l/\sigma_k^2}{\sum_{j=l}^{K} l/\sigma_j^2} \quad (9)$$

This weighting can significantly suppress the influence of low-quality sources when the variance of noise between sources is large.

Median fusion takes the median of each source principal variable as the fusion value at each time step. Compared with the mean, the median is not affected by extreme values and has natural robustness to residual anomalies in individual sources. It does not require accurate estimation of the variance of each source and is suitable for scenarios where anomaly detection is incomplete and there are still a small number of outliers, but it may lose some of the denoising gain of multi-source averages.

Principal component analysis fusion first standardizes the principal variables of each source, and then performs feature decomposition on the covariance matrix [10-11]. The mathematical expression for feature decomposition is shown in formula (10).

$$\Sigma = UAU^T \quad (10)$$

Take the main component corresponding to the maximum characteristic value as the fusion direction, and map the score linearly back to the true value scale. The direction of the principal component is automatically determined by the data variance structure, which is equivalent to finding the optimal linear combination of each source observation in the least squares sense.

The four methods each have conditions: equal weighted average demand source quality is close; The reciprocal weighting of variance requires that the variance is estimable and the error is independent; Median fusion anti residual anomaly; Principal component fusion is easily dominated by large and diverse sources. They are all static and do not use timing; Kalman filtering can model dynamic systems, recursively fuse multiple sources, and obtain the minimum variance estimate.

5. Experimental Analysis

5.1. Experimental Evaluation Metrics

The experimental data consists of simulated observations from a three source information network, with a duration of 2000 observation points. The dirty data injected by each source includes random missing data with missing rates of 10%, 15%, and 20%, respectively; 5% of abnormal observations have abnormal amplitudes ranging from several times to more than ten times the standard deviation of normal noise. The experimental evaluation indicators of this article are root mean square error (RMSE) and mean absolute error (MAE); Accuracy, recall, and F1 score of anomaly detection; The accuracy of three classification state determination.

Abnormal detection use TP, FP and FN to represent real cases, false cases and false negative cases respectively, and the accuracy rate, recall rate and F1 are defined as formulas (11).

$$P = \frac{TP}{TP+FP}, R = \frac{TP}{TP+FN}, F_1 = \frac{2PR}{P+R} \quad (11)$$

The proportion of samples with three classification accuracy statistics that are consistent with the predicted state and the true state. The experiment was conducted in a Python environment with a fixed random seed to ensure reproducibility of the results.

5.2. Data Cleaning Effect Analysis

Table 1 shows the comparison of quality indicators before and after data cleaning. In terms of integrity, it has been improved from 90.07% to 100%, and the insertion link has completely eliminated the lack. In terms of accuracy, measured by MAE, it decreased from 2.174 to 1.031, a decrease of 52.6%. The quality ratio before and after cleaning is shown in Table 1.

Table 1. Comparison of Data Quality Indicators Before and After Cleaning

Completeness (front → back)	Accuracy (front → back)	Consistency (front → back)	Timeliness (from earlier to later)	MAE (front → back)
0.9007 → 1.0000	0.9598 → 0.9809	1.0000→1.0000	1.0000→1.0000	2.174 → 1.031

6. Conclusion and Outlook

This article is aimed at the multi-source data quality management of the information network, and puts forward a cleaning and integration framework with feature engineering as the main line. Establish a four-dimensional quality evaluation of integrity, accuracy, consistency and timeliness in modeling. There are five links of cleaning design time alignment, missing interpolation, abnormal detection and repair, duplicate identification, and source-level deviation calibration, using EM interpolation, de-trend 3σ detection, de-weighting and calibration. The inverse weighting of variance and principal component analysis are used at the fusion feature level, and majority voting and D-S evidence are used at the decision-making level to achieve multi-level utilization. Follow-up work: First, introduce real data sets and typical faults to verify the impact of noise correlation, pulse interference and large-scale missing; Second, introduce inter-source correlation into interpolation and fusion, add covariance structured constraints in EM, and use generalized least squares weighting for fusion; Third, study incremental cleaning and online fusion, so as to card Erman filter recursion is used as a reference to achieve dynamic update, reduce delay and adapt to distribution drift.

References

[1] Zhang, L. (2026). Research on data cleaning of highway control point re-measurement based on deep long short-term memory network. China Water Transport, (18), 118–120. https://doi.org/10.13646/j.cnki.42-1395/u.2026.18.042

[2] Liu, Z. C., Ye, Y. Z., Xie, P. Y., et al. (2026). A method for cleaning flood reporting data by integrating unsupervised learning and attention framework. People's Yangtze River. Advance online publication. https://link.cnki.net/urlid/42.1202.TV.20260820.1817.010

[3] Zheng, Q., & He, Y. Q. (2026). A method for power engineering information analysis based on multi-source data feature extraction. Electronic Design Engineering, 34(11), 152–156. https://doi.org/10.14022/j.issn1674-6236.2026.11.029

[4] Shi, X. M., Li, R., & Xia, Y. L. (2026). Adaptive fusion method for distribution network engineering data based on heterogeneous feature deep extraction algorithm. Microcomputer Applications, 42(6), 248–251, 255.

[5] Feng, Q. H. (2026). Aircraft target recognition method based on EM algorithm. China Science and Technology Information, 38(9), 73–75.

[6] Zhao, Z. W., Yuan, Y., & Yang, K. (2026). Parameter estimation of power-uniform mixture distribution based on EM algorithm. Journal of Shenyang Normal University (Natural Science Edition), 44(1), 81–89.

[7] Shen, Y. J., Jin, Y. Y., & Zhang, H. Z. (2026). Parameter estimation for ARMA(p,q) linear measurement error model based on EM algorithm. Journal of Hangzhou Normal University (Natural Sciences Edition), 25(2), 206–215. https://doi.org/10.19926/j.cnki.issn.1674-232X.2024.10.081

[8] Zhang, Q. Y., Wang, D. H., Zhang, R. G., et al. (2025). A method for outlier processing of fragment velocity measurement based on the 3σ criterion. Shipboard Electronic Engineering, 45(10), 169–172, 184.

[9] Xin, G. H., Li, Z. A., Lin, D. C., et al. (2025). Weak target detection in infrared thermal imaging based on random gray-level fluctuation field and 3σ criterion. Journal of Jinan University (Natural Science Edition), 39(5), 706–718. https://doi.org/10.13349/j.cnki.jdxbn.20250822.001

[10] Yang, G., Cheng, L., Xu, W. Y., et al. (2026). Bearing fault diagnosis algorithm based on multi-objective optimization feature mode decomposition. Bearing, (8), 101–109. https://doi.org/10.19533/j.issn1000-3762.202401029

[11] Ye, W., & Deng, J. F. (2026). A parking occupancy prediction method based on feature decomposition and meta-learning. Times Automobile, (16), 13–15.