A Feature Extraction Method Based on Feature Fusion and its Application in the Text-Driven Failure Diagnosis Field

Shenghan Zhou, Bang Chen, Yue Zhang, HouXiang Liu, Yiyong Xiao*, Xing Pan

School of Reliability and Systems Engineering, Beihang University, Beijing 100191 (China)

Received 20 June 2020 | Accepted 28 October 2020 | Published 11 November 2020

Abstract

As a basic task in NLP (Natural Language Processing), feature extraction directly determines the quality of text clustering and text classification. However, the commonly used TF-IDF (Term Frequency & Inverse Document Frequency) and LDA (Latent Dirichlet Allocation) text feature extraction methods have shortcomings in not considering the text's context and blindness to the topic of the corpus. This study builds a feature extraction algorithm and application scenarios in the field of failure diagnosis. A text-driven failure diagnosis model is designed to classify and automatically judge which failure mode the failure described in the text belongs to once a failure-description text is entered. To verify the effectiveness of the proposed feature extraction algorithm and failure diagnosis model, a long-term accumulated failure description text of an aircraft maintenance and support system was used as a subject to conduct an empirical study. The final experimental results also show that the proposed feature extraction method can effectively improve the effect of clustering, and the proposed failure diagnosis model achieves high accuracies and low false alarm rates.

Keywords

Feature Extraction, Feature Fusion, Text-Driven, Failure Diagnosis.

DOI: 10.9781/ijimai.2020.11.006

I. Introduction

IN the information age, data is generated all the time, especially text data. Because of its convenience, flexibility and universality, the quantity of text data is growing exponentially in, for example, applications such as twitter, online articles and shopping reviews. These words usually contain a lot of useful information, but also a lot of interference information. So, the text data needs to be processed to mine the useful information.

For structured data, scholars usually write corresponding programs directly to mine information with the help of computers. However, for unstructured data such as text, the computer cannot recognize it directly. Therefore, it is necessary for the computer to first understand the text before mining the text for information. The recognition process involves converting text symbols into numeric symbols that can be recognized by computers. Scholars have tried many methods to resolve issues around recognition, but the original rule-based methods obviously do not solve this problem very well [1]. However, inspired by neural network, many scholars put forward some effective new technologies [2], and text feature extraction is one of them.

Since computer cannot directly recognize unstructured data, the text needs to be transformed into a structured format for processing. At present, the VSM (Vector Space Model) [3] is widely used to make the transformation; but a document may contain thousands of words, which easily leads to dimension explosion and expensive calculations. Therefore, feature extraction is usually used for further dimension reduction. Feature extraction can find the most representative text features that use low-dimensional feature vectors to represent text data based on the original text [4]. TF-IDF [5] and LAD [6], are two classical models for text feature extraction, and are easy to operate. However, TF-IDF only extracts keywords without considering the context. Although LDA considers the context, it tends to lose features and be ambiguous. Inspired by Zhao [7] et al. 's work of fusing χ^2 statistics-based feature and LDA semantic-based feature to improve the performance of a feature extraction, this paper proposes a text feature extraction method based on feature fusion, which combines the TF-IDF and LDA methods.

As NPL technology advanced, it was used in the field of text-driven failure diagnosis. Dnyanesh [8] et al. proposed a novel ontology-based text mining methodology to construct the D-matrices by automatically mining the unstructured repair verbatim data collected during failure diagnosis and used it to do failure pattern recognition. Rodrigues [9] et al. used text mining and neural networks to identify and classify aircraft failure patterns. However, most of the current text-based failure diagnosis models are supervised, which means these models apply to only text data with labels. However, in fact, the failure-description text is often unlabeled because of high labor costs. Therefore, based on the proposed feature extraction method, this paper developed a text-driven unsupervised failure diagnosis model. The extracted feature vectors are clustered to obtain the pseudolabel data, and then the pseudolabel data is used to train the classifier for feature diagnosis. This failure diagnosis model can classify and automatically judge which failure mode the failure described in the text belongs to once a failure-description text is entered.

* Corresponding author.
E-mail address: xiaoyiyong@buaa.edu.cn

To verify the effectiveness of the proposed feature extraction algorithm and failure diagnosis model, a long-term accumulated failure-description text of an aircraft maintenance and support system was used as a subject to conduct an empirical study.

The main contribution of this paper is to propose a more effective feature extraction method, by fusing TF-IDF and LDA, two typical feature extraction methods, and apply it to the field of failure diagnosis, by establishing an intelligent text-driven failure diagnosis model with the help of machine learning methods such as clustering and classification. The remaining of this paper is organized as follows: Section II presents an overview of text feature extraction and feature fusion. Section III is the introduction of the basic principle of the proposed feature extraction algorithm. Section IV presents the proposed text-driven failure diagnosis model. Section V is the experiment and the discussion of the experimental results. Section VI covers conclusions and discussion.

II. Related Work

A. Text Feature Extraction

As the basic work of text processing, text feature extraction has always been a hot research topic in NLP. So far, most of the existing feature extraction methods are based on the bag-of-words model [10], the topic model [11] and the word embedding model [12] - [13].

The bag-of-words model [10] adopts one-hot encoding to generate word vectors. Each word vector's dimension is equal to the size of the word vocabulary. In this vector, only one dimension's value is 1 and the rest are 0. Obviously, this kind of vector composed of 0 and 1, cannot represent a word accurately, because different words have a different importance to the text. Scholars usually use the TF-IDF method to assign weights to one-hot vectors. TF-IDF [5] is a widely used weighting technique, and plays an important role in the field of information retrieval. A TF-IDF is easy to carry out and usually performs well in short-text dataset, but performs badly in a long-text dataset or a class imbalance dataset.

A topic model [11] is a kind of topic generation models and a three-layer Bayesian probability model. The topic model's core idea is that a document selects a topic according to a certain probability, and a topic also selects a word according to a certain probability. LDA [6], as a classical topic model, is widely used in the task of text classification and text clustering. Compared with TF-IDF, LDA considers the context and performs much better in long-text dataset. However, LDA, as an unsupervised model, tends to lose features, is ambiguous in the process of feature extraction and performs badly in short-text datasets.

A word embedding model is designed to solve the dimension explosion problem of the bag-of-words model when processing long-text dataset. Through neural networks [13], word cooccurrence matrices, probabilistic models and other methods, a word embedding model maps the bag-of-words model's one-hot vector to a continuous vector space with a much lower dimension to enable a dimension reduction [12] - [13]. Word2vec is the most widely used word embedding model framework; it includes two word-vector-generation models, Skip-Gram and CBOW (Continuous Bag-of-Words Model). Skip-Gram and CBOW are three-layer neural networks with different inputs. Skip-Gram inputs the current word to predict the surrounding words, while CBOW inputs the surrounding words to predict the current word. Obviously, the two models take the context into account. However, the two models usually perform badly in short-text datasets [14].

B. Feature Fusion

Feature fusion originates from data fusions that were originally conducted in the military. In recent years, with the rapid development of AI, data fusion has been widely applied in intelligent medical [15], intelligent industry [16], intelligent transportation [17] and so on. Data fusion is a framework, which contains fusion modes and tools. Data fusion mainly uses different fusion modes and tools to combine different data sources, which may generate improved new data for certain application scenarios [18]. Whether the data after a fusion is effective or not mainly depends on the application scenario. In most cases, data fusion can effectively enhance the authenticity and availability of data [19], which is why data fusion is needed.

Feature fusion, as a technology of data fusion; uses given feature sets to generate new fusion features [19], and is very suitable for classification tasks. Liu [20] et al. fused two groups of feature vectors into a unit vector, and extracted features from high-dimensional vector space. They proposed a serial feature fusion algorithm and applied this algorithm to do face recognition. Their experiments showed that this algorithm could reach an accuracy rate of 98.5% with only 25 features. Yang [19] et al. proposed a new serial feature fusion algorithm for unstructured data, and tested it on the CENPARMI handwritten digital library, the NUST603 handwritten Chinese character library and the ORL face image library. The experimental results showed that their algorithm effectively improved the classification accuracy. Sun [21] et al. also proposed a new feature fusion algorithm based on CCA (Canonical Correlation Analysis), which performed well in small sample dataset with high dimensions. They first extracted two groups of feature vectors with the same pattern, then established correlation criterion functions between them, and finally extracted their representative features to form effective recognition vectors.

III. Text Feature Extraction Method Based on Feature Fusion

Text feature extraction is a process of text vectorization. As the first step of text processing, it directly determines the effects of the follow-up processes. However, text feature extraction is a very complex problem, because it involves conversion of abstract character symbols into concrete number symbols under the premise of maintaining the meaning of the original text. TF-IDF and LDA are two commonly used feature extraction methods proposed by scholars. The TF-IDF method is easy to execute and usually performs well in short-text datasets, but ignores the context. LDA considers the context and performs well in long-text dataset, but easily leads to blindness. Therefore, this paper proposes a text feature extraction method based on feature fusion, which combines the TF-IDF and LDA methods and is named TI-LDA.

A. TF-IDF Feature Extraction Method

The TF-IDF feature extraction method is actually the TF-IDF weighting of one-hot vector generated by a bag-of-words model. A One-hot vector whose dimension is equal to the size of the word vocabulary is discrete, and only one dimension's value is 1 and the rest are 0. The one-hot representation of a sentence can be obtained by adding the one-hot vector of all the words in the sentence. However, obviously, not every word is equally important to a sentence, so using TF-IDF to weight each word is necessary. The core idea of TF-IDF is the importance of a word is positively correlated with the frequency of its occurrence in a given text, and negatively correlated with the frequency of its occurrence in all texts of the corpus. The TF-IDF is composed of TF (Term Frequency) and IDF (Inverse Document Frequency). TF refers to the frequency of a word's occurrence in a given text, and its calculation formula is:

$$\mathrm{TF}_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \tag{1}$$

where $n_{i,j}$ is the number of times that word t_i appears in text d_j and

the denominator is the total number of times that all word appears in text d_j.

IDF is used to measure the frequency of a word's occurrence in all texts of the corpus, and its calculation formula is:

$$IDF_i = log_2 \frac{|D|}{|\{j : t_i \in d_j\}|+1} \tag{2}$$

where $|D|$ is the total number of texts in the corpus and $|\{j : t_i \in d_j\}|$ is the number of texts containing t_i in the corpus.

The calculation formula of TF-IDF is:

$$TF_{i,j}IDF_i = TF_{i,j} \times IDF_i \tag{3}$$

The normalized formula of TF-IDF is:

$$TF_{i,j}IDF_i' = \frac{TF_{i,j}IDF_i}{\sqrt{\sum_{j=1}^{|d_j|}(TF_{i,j}IDF_i)^2}} \tag{4}$$

The TF-IDF algorithm is simple in principle, easy to operate and efficient to calculate, which make it suitable for short-text mining. However, it ignores the context to carry out vectorization and easily causes dimension explosion when dealing with long text.

B. LDA Feature Extraction Method

LDA is a statistical topic model which represents the topics of each document in the form of a probability distribution. An LDA believes that a document consists of several topics, and each topic consists of several words. When generating a document d with K topics, the probability of a word w being selected is:

$$p(w|d) = \sum_K p(w|t_k) \times p(t_k|d) \tag{5}$$

where K and k are the total number of topics and indexes of the group of topics and t_k stands for the topic k. For example, there are three topics: animals, actions and names. Different words are distributed under each topic. "cat", "dog" and "pig" belong to the topic of animals. "sitting", "running" and "standing" belong to the topic of actions. "Tony", "Jack" and "Lucy" belong to the topic of names. Suppose a sentence that says the dog is sitting need to be generated. The first step is to select the topics under the condition of the target semantics. The second step is to choose the words under the condition of the selected topics. In the end, the sentence "the dog is sitting" can be generated.

Therefore, the distribution of words in a given document can be obtained as Fig. 1.

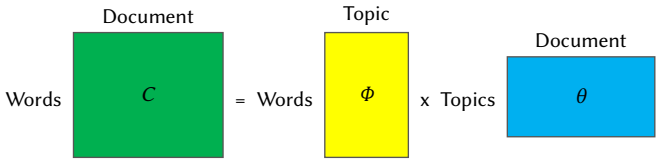

Fig. 1. LDA Model Diagram.

The process of generating a document by the LDA model can be summarized as detailed below:

- Step 1: Sample from the Dirichlet distribution α to generate the topic distribution θ_i of document i.;
- Step 2: Sample from the Multinomial distribution θ_i to obtain the topic $z_{i,j}$ of word j in document i.;
- Step 3: Sample from the Dirichlet distribution β to generate the topic distribution $\phi_{z_{i,j}}$ of topic $z_{i,j}$;
- Step 4: Sample from the Multinomial distribution $\phi_{z_{i,j}}$ to obtain word $\omega_{i,j}$.

Based on the above process, the following joint distribution can be obtained:

$$p(\omega_i, z_i, \theta_i, \Phi|\alpha, \beta) = \prod_{j=1}^{N} p(\theta_i|\alpha)p(z_{i,j}|\theta_i)p(\Phi|\beta)\, p\left(\omega_{i,j}\left|\theta_{z_{i,j}}\right.\right) \tag{6}$$

By integrating θ_i and Φ, and summing z_i, the maximum likelihood estimation of the word distribution can be obtained as follows:

$$p(\omega_i|\alpha, \beta) = \int_{\theta_i}\int_{\Phi}\sum_{z_i} p(\omega_i, z_i, \theta_i, \Phi|\alpha,\beta) \tag{7}$$

Finally, parameters in an LDA model can be obtained by a Gibbs sampling [22].

The LDA model can effectively extract the semantic information of the text, in consideration of the context. However, in the selection of topic words, LDA has a certain blindness, that easily causes ambiguity and feature loss.

C. TI-LDA Feature Extraction Method

The feature fusion algorithm has been widely used in the field of AI for applications such as target tracking, pattern recognition and image understanding. In the field of pattern recognition, Jian Yang [19] et al. proposed two fusion strategies, parallel feature fusion and serial feature fusion. They also verified the robustness and practicability of the two fusion strategies through experimentation.

Suppose A and B are two different feature spaces of sample space Ω, meanwhile suppose $\alpha \in R^n$ and $\beta \in R^n$ are a feature vector of A and B, respectively. Then, the parallel feature fusion can be expressed as:

$$\gamma = \alpha + i\beta \tag{8}$$

where i is the imaginary component and $\gamma \in R^{\max(n,m)}$ is a feature vector of the new feature spaces. In the parallel feature fusion process, feature vectors α and β may have different dimensions; the feature vector with the lower dimension needs to be supplemented with 0 before fusion. Take $\alpha = (a_1, a_2, a_3)^T$ and $\beta = (b_1, b_2)^T$ for example. First, add 0 to supplement vector β to create a three-dimensional feature vector $(b_1, b_2, 0)^T$, then carry out the feature fusion according to equation (8), and the final result is $\gamma = (a_1 + ib_1, a_2 + ib_2, a_3 + i\,0)^T$.

Serial feature fusion can be expressed as:

$$\gamma = \begin{pmatrix}\alpha \\ \beta\end{pmatrix} \tag{9}$$

Compared with parallel feature fusion, serial feature fusion does not need to consider the vectors with different dimension. Also take $\alpha = (a_1, a_2, a_3)^T$ and $\beta = (b_1, b_2)^T$ for example; fuse directly and the final result is $\gamma = (a_1, a_2, a_3, b_1, b_2)^T$.

For a given text sample space Ω, suppose A and B is the feature vector space based on TF-IDF and LDA, meanwhile $\alpha_i \in R^n$ is a feature vector in A and $\beta_i \in R^m$ is a feature vector in B. Adopting the parallel feature fusion strategy, according to equations (8), the sample space Ω can be represented as:

$$\Omega = (\gamma_1, \gamma_2, \ldots, \gamma_j)^T = \begin{vmatrix}\gamma_{11} & \gamma_{12} & \cdots & \gamma_{1k} \\ \gamma_{21} & \gamma_{22} & \cdots & \gamma_{2k} \\ \cdots & \cdots & \cdots & \cdots \\ \gamma_{j1} & \gamma_{j2} & \cdots & \gamma_{jk}\end{vmatrix} = \begin{vmatrix}\alpha_{11} + i\beta_{11} & \alpha_{12} + \beta_{12} & \cdots & \alpha_{1k} + \beta_{1k} \\ \alpha_{21} + i\beta_{21} & \alpha_{22} + i\beta_{22} & \cdots & \alpha_{2k} + i\beta_{2k} \\ \cdots & \cdots & \cdots & \cdots \\ \alpha_{j1} + i\beta_{j1} & \alpha_{j2} + i\beta_{j2} & \cdots & \alpha_{jk} + i\beta_{jk}\end{vmatrix} \tag{10}$$

While adopting the serial feature fusion strategy, according to equations (9), the sample space Ω can be represented as:

$$\Omega = (\gamma_1, \gamma_2, \ldots, \gamma_j)^T = \begin{vmatrix} \gamma_{11} & \gamma_{12} & \cdots & \gamma_{1k} \\ \gamma_{21} & \gamma_{22} & \cdots & \gamma_{2k} \\ \cdots & \cdots & \cdots & \cdots \\ \gamma_{j1} & \gamma_{j2} & \cdots & \gamma_{jk} \end{vmatrix} = \begin{vmatrix} \alpha_{11} & \alpha_{12} & \cdots & \alpha_{1k} \\ \beta_{11} & \beta_{12} & \cdots & \beta_{1k} \\ \alpha_{21} & \alpha_{22} & \cdots & \alpha_{2k} \\ \beta_{21} & \beta_{22} & \cdots & \beta_{2k} \\ \cdots & \cdots & \cdots & \cdots \\ \alpha_{j1} & \alpha_{j2} & \cdots & \alpha_{jk} \\ \beta_{j1} & \beta_{j2} & \cdots & \beta_{jk} \end{vmatrix} \quad (11)$$

IV. Text-Driven Failure Diagnosis Model Based on TI-LDA

To make full use of failure-description text and understand the role the TI-LDA text feature extraction method plays in the field of fault diagnosis, this paper researched TI-LDA in the failure diagnosis field and designed a text-driven failure diagnosis model, that is suitable for small data samples and the main framework is as shown in Fig. 2. Before any further processing, the text data needs to be preprocessed. Specifically, for English text data, stop words need to be removed, while for Chinese text, word segmentation is also needed because there is no distinct identifier for separation. In addition, there are differences in word granularity, part of speech, polyphonic characters and so on between Chinese NLP and English NLP. Although the model is mainly for Chinese text, there is usually a mixed use of Chinese and English for failure-description text. For this situation, the model treats English words in the text as special Chinese characters. After the preprocessing, feature extraction is done to obtain text vectors, the obtained feature vectors are processed by CFSFDP (Clustering by Fast Search and Find of Density Peaks) clustering to mark the pseudolabels for failure text. The obtained pseudolabel data cannot be directly put into the classifier for training, because the class imbalance problem often exists in the failure text, which will affect the performance of the classifier. Therefore, this paper adopts the SMOTE (Synthetic Minority Oversampling Technique) oversampling method to balance the pseudolabel data. Finally, the balanced data is put into the SVM classifier to train the failure diagnosis model.

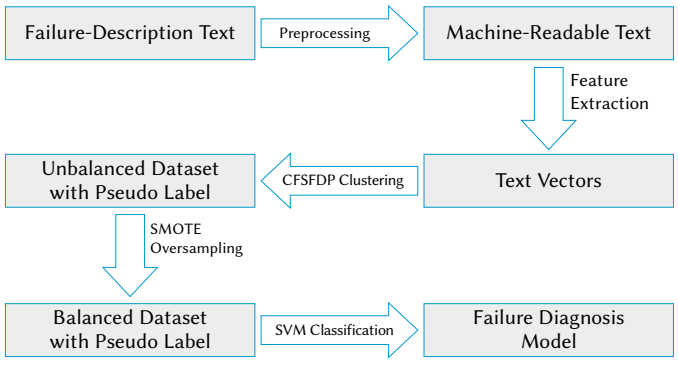

Fig. 2. Flow Chart of Aircraft Failure Diagnosis Model.

A. Text Preprocessing

As mentioned above, different preprocessing strategies should be adopted for English text and Chinese text. The English text should be processed by removing stop words directly, while the Chinese text should be segmented first and then the stop words should be removed. At present, the common Chinese word segmentation methods are mainly based on dictionaries, statistics and rules, and a dictionary-based method is the most effective and widely used. Common dictionary-based word segmentation systems include Jieba, the CAS (Chinese Academy of Sciences) segmentation system, Smallseg and Snailseg. Their functions are compared in Table I. It can be seen from Table I that Jieba is more powerful and suitable for the text data used in this article, so this paper adopted Jieba for word segmentation.

TABLE I. Comparison of Dictionary-Based Word Segmentation Systems

Segmentation system	Custom dictionary	Part-of-speech tagging	Keyword extraction
Jieba	√	√	√
CAS	√	√	×
Smallseg	√	×	×
Snailseg	×	×	×

There are often a large number of stop words in the text, such as emotional particles and punctuation marks, that have no contribution to the semantic expression. If text is directly used for subsequent processing, these stop words will inevitably cause too high a dimension for the text vector, increase the calculation cost, and interfere with text clustering. Therefore, these stop words need to be removed with the help of a stop words list. A simpler approach is to directly use the stop word list established by professional organizations, such as the NLTK (Natural Language Toolkit) English stop word list and the Baidu Chinese stop word list. Although this method is simple, it does not achieve the best effect. If you want to achieve the best effect, you need to establish a special stop table in accordance with the specific situation of the text data.

B. Feature Extraction

The text feature extraction function adopts the TI-LDA method proposed in this paper. In terms of the selection of fusion strategy, this paper selects the serial pattern feature fusion. By comparing equations (10) and (11), it can be found that the parallel feature fusion will continue to reduce the dimension. However, before the fusion of text features, the preprocessing and feature extraction will have already reduced the dimension of the text data. Obviously, more features will be lost if we use the parallel feature fusion. Therefore, this paper uses the serial feature fusion to do feature fusion.

C. Text Clustering Based on CFSFDP

Text clustering is a key element in the failure diagnosis model, and its main function is to mark the pseudolabels for the failure text. Therefore, choosing a clustering method suitable for the text data is very important. The current clustering algorithms can be broadly divided into partition-based methods [23], hierarchical-based methods [24] - [25], density-based methods [26], grid-based methods [27] and model-based methods. Because the failure-description text studied in this paper is typical of small data sampls, this paper uses CFSFDP to do the clustering. CFSFDP is a clustering method for small data sampls published in Science by Rodriguez [28] et al. Compared with the typical partition-based K-Means [29] method, CFSFDP not only handles clusters with aspherical shapes but also automatically determines the number of clusters. Compared with the typical density-based DBSCN (Density-Based Spatial Clustering of Applications with Noise) method, CFSFDP doesn't need to iterate repeatedly to determine the density threshold.

CFSFDP assumes that the center of the cluster is surrounded by some points with low local density, and these points are far away from other points with high local density. Therefore, the clustering centers can be obtained by calculating the nearest distance, and the remaining points can be divided into their categories according to their order of density. Suppose p_i and p_j are two different points of discrete data point

set D = $\{p_1, p_2, ..., p_n\}$, and define p_i's local density ρ_i as the number of points in the circle with p_i as the center and d_c as the radius. Then, ρ_i can be calculated by the following formulas:

$$\rho_i = \sum_j \chi(d_{ij} - d_c) \tag{12}$$

where function:

$$\chi(x) = \begin{cases} 1, x < 0 \\ 0, x \geq 0 \end{cases} \tag{13}$$

Here, d_{ij} is the distance between p_i and p_j, and d_c is the cutoff distance, that needs to be determined manually.

Define the set of points with a higher density than p_i as $I_S^i = \{k \in I_S : \rho_k > \rho_i\}$, define the distance δ_i to be:

$$\delta_i = \begin{cases} \min\limits_{j \in I_S^i} d_{ij} & I_S^i \neq \emptyset \\ \max\limits_{j \in I_S} d_{ij} & I_S^i = \emptyset \end{cases} \tag{14}$$

When the data point p_i has the largest local density, $I_S^i = \emptyset$ and δ_i represents the maximum distance between p_i and p_j in the data set I_S, otherwise δ_i represents the minimum distance between p_i and p_j in the data set I_S^i.

To comprehensively measure the local density p_i and distance δ_i, another variable γ_i needs to be introduced. In addition, the calculation criteria of γ_i is:

$$\gamma_i = \rho_i \delta_i \tag{15}$$

The clustering centers can be selected according to the value of γ_i, because the clustering centers usually have a larger value of γ_i. If all values of γ_i are arranged in descending order and plotted on a two-dimensional plane, it can be found that the values of a in the nonclustering central interval are relatively smooth. If we arrange all γ_i in descending order and plot them in coordinates, you can see the value of γ_i is generally small and changes stably in the interval of non-nonclustering centers, while the value of γ_i is generally large in the interval of clustering centers and there is an obvious jump of γ_i's value near the critical point. Therefore, the number of clustering centers and classes can be determined based on the above characteristics.

D. Oversampling Based on SMOTE

In most of the failure monitoring data, including the failure-description text, there exists a class imbalance problem. The major class usually has more samples than the minor class. In fact, the major class occurs frequently but usually does less harm, while the minor class occurs occasionally but does great harm. This kind of unbalanced class data is a great challenge to classification. If the unbalanced data is directly used for classification, the minor-class samples will be submerged in the major class samples, which often results in high false alarm rates for the major-class and high missing alarm rates for the minor-class [30].

At present, there are mainly two ways of data balancing processing of oversampling. One is directly copying the samples of the minor class, the other is artificially generating the samples of the minor class according to the minor class's characteristics. The former is easy to do but easily causes overfitting, while the latter is more complex but difficult to overfit. This paper adopts the SMOTE [31] oversampling method, which is based on the latter, because the sample size of the data used in this paper is small.

Based on the above considerations, the SMOTE [31] algorithm, a widely used and relatively mature oversampling method, was adopted in this paper. The basic function of SMOTE is to manually add the minor-class samples to the new sample set by analyzing the characteristics of the minor-class. The SMOTE processes to solve a class imbalanced problem are as follows:

Define a sample set $X = \{x_i | i = 1,2,3, \dots, m\}$ of a minor class. For any point in X, calculate the Euclidean distance between this point and all remaining points to obtain the k nearest points. Here, this paper assumes the multiplier of oversampling as n, that is randomly selecting n points in the k nearest points to generate set $Y = \{\hat{x}_j | j = 1,2,3, \dots, n\}$. By a random linear interpolation, add the new sample to X, which is shown in the following formula:

$$x_{new}(i,j) = \{x_i + rand(0,1)(\hat{x}_j - x_i) | i = 1,2,3, \dots, m; j = 1,2,3, \dots, n\} \tag{16}$$

where $rand(0,1)$ is a random number between 0 and 1. The above formula can generate m samples of the minor class to achieve the purpose of balancing the data set.

E. Classification Based on SVM

The pseudo-label data after balanced processing, needs to be put into the classifier for training. To do the selection of the classifier, the SVM classifier is selected in this paper. SVM, as a supervised learning method suitable for data with small sample sizes, was first proposed by Vapnik et al. [32] Due to its characteristics of easy operation and high robustness, SVM has been widely used in the field of feature diagnosis, and this paper also uses SVM to do classification. SVM is mainly based on statistics. SVM first maps the input data from the low-dimensional space to the high-dimensional space to make the problem linearly separable, then finds an optimal hyperplane in the high-dimensional space to divide the data. Therefore, the selection of the optimal hyperplane directly determines the classification effect.

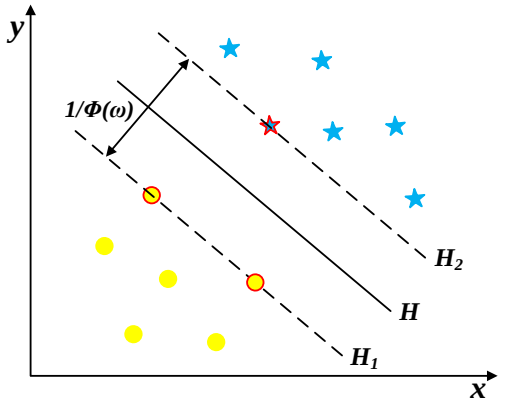

Fig. 3. Linear Binary Classification Diagram.

Because the linear binary classification of SVM is the basis and prototype of SVM. This process first consider the linear binary classification problem, as shown in Fig. 3. Define the sample points (x_i, y_i), $i = 1, 2, ..., s$, $x_i \in R^m$, $y_i \in \{1, -1\}$. Based on the above conditions, a classification hyperplane H is constructed:

$$wx + b = 0 \tag{17}$$

H can divide the above sample points into two classes, and the formula is:

$$\begin{cases} wx_i + b \geq 0 & y_i = 1 \\ wx_i + b \leq 0 & y_i = -1 \end{cases} \tag{18}$$

H can separate two different classes of samples, but the goal of SVM is to find the optimal hyperplane, that is, the maximum distance between the two classes of samples. Therefore, the objective function is:

$$min\Phi(w) = ||w||^2/2 \tag{19}$$

The constraint is:

$$y(wx_i + b) - 1 \geq 0 \quad i = 1,2,\dots,l \tag{20}$$

To find the optimal solution of equation (19), the Lagrange multiplier is introduced, and equation (19) can be linearized into

$$\min_a W(a) = \sum_{i=1}^{l} a_i - \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l} a_i a_j y_i y_j < x_i \bullet x_j > \tag{21}$$

where $a_i \geq 0$ and $i = 1, 2, ..., l$.

The constraint is updated to:

$$\sum_{i=1}^{l} y_i a_i = 0, a_i \geq 0 \tag{22}$$

When $a_i \geq 0$, these sample points are referred to as support vectors. The optimal classification discriminant function is:

$$f(x) = sgn(\sum y_i a_i (x \bullet x_i) + b) \tag{23}$$

So far, the linear binary classification problem is solved. In addition, the nonlinear binary classification problem can be converted to a linear binary classification problem by kernel functions. The objective function of the nonlinear binary classification problem is:

$$\min_a W(a) = \sum_{i=1}^{l} a_i - \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l} a_i a_j y_i y_j K < x_i \bullet x_j > \tag{24}$$

The optimal classification discriminant function of the nonlinear binary classification problem is:

$$f(x) = sgn(\sum y_i a_i K(x \bullet x_i) + b) \tag{25}$$

Here, $K(x \bullet x_i)$ is the kernel. The kernel function needs to be selected artificially, and the Gauss kernel function is used in this paper:

$$K(||x_i - x_c||) = exp(\frac{-||x_i - x_c||^2}{2\sigma^2}) \tag{26}$$

where x_c is the center of the kernel function, and σ is the width parameter of the kernel function.

So far, the binary classification problem has been solved. For the multiclassification problem, this paper builds an SVM multiclassification framework based on a binary tree, which is shown in Fig. 4.

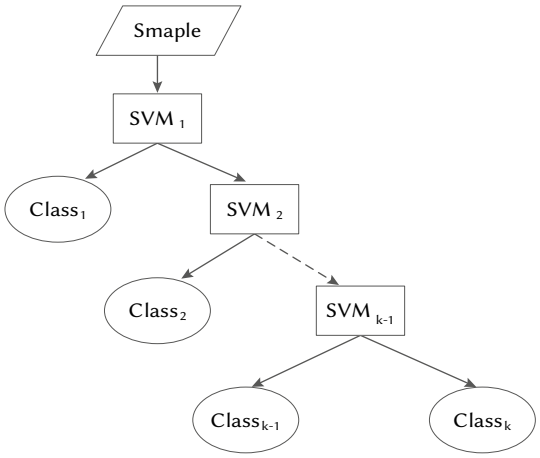

Fig. 4. The SVM Multiclassification Framework.

V. Experiments and Result Analysis

To verify the effectiveness of the proposed TI-LDA feature extraction algorithm and the text-driven failure diagnosis model, this paper used the failure-description Chinese text accumulated and recorded by an aircraft maintenance and support system to design verification experiments and analyzes the experimental results in detail. After eliminating the repeated and missing data, 1683 effective failure-description texts are obtained. In addition, some failure-description texts are shown in Table II.

TABLE II. Examples of Aircraft Failure-Description Text

Text number	Content
1	电机不工作，离合器坏 Motor didn't work, and clutch was damaged
2	渗油，密封圈损坏 Oil leakage, and seal ring was damaged
3	漏油致起落架放不下，密封圈损坏 Oil leakage caused undercarriage unable to be put down, and seal ring was damaged
…	…
1682	四台发动机尾喷口堵盖过大 The block cover of four engines' jet nozzle was too large
1683	信号灯故障 Signal lamp had a breakdown

To obtain better effectiveness for removing the stop words, this paper designed a special stop words list according to the characteristics of the corpus and the existing stop word list. Through the analysis of the text used in this paper, it is easy to find that two-word stop words are the most common, followed by professional characters, letters, and three-word stop words. Therefore, based on the above characteristics and the commonly used stop words lists, this paper designed a special stop word list to removing stop words. Some of the stop words are shown in Table III.

TABLE III. Stop Words List

Number	Stop word	Number	Stop word
1	———	8	]
2	》），	9	?
3	）÷（1 -	10	。
4	”，	11	竟然 (Actually)
5	℃	12	看 (Look)
6	[⑤]]	13	快 (Almost)
7	[	…	…

A. Effectiveness Verification Experiment of TI-LDA

To verify the effectiveness of TI-LDA, this paper first used TF-IDF, LDA and TI-LDA to extract the feature vectors of the failure-description texts respectively, then clustered the three sets of feature vectors using CFSFDP, and finally evaluated the effectiveness of the feature extraction by comparing the effects of clustering.

This paper used TF-IDF, LDA and TI-LDA to extract the features of the preprocessed text, and the normalized feature vectors obtained are shown in Table IV, Table V and Table VI.

TABLE IV. Normalized Feature Vectors of TF-IDF

Number	Dimension					
	1	2	3	…	2310	2311
1	0.662	0.350	0.000	…	0.000	0.000
2	0.000	0.350	0.270	…	0.000	0.000
3	0.662	0.000	0.270	…	0.000	0.000
…	…	…	…	…	…	…
1682	0.000	0.000	0.000	…	0.781	0.000
1683	0.000	0.000	0.000	…	0.000	0.599

TABLE V. Normalized Feature Vectors of LDA

Number	Dimension					
	1	2	3	...	2310	2311
1	0.726	0.056	0.056	...	0.000	0.000
2	0.042	0.042	0.043	...	0.000	0.000
3	0.000	0.971	0.000	...	0.000	0.000
...	...	...	...	...	...	...
1682	0.913	0.019	0.420	...	0.000	0.000
1683	0.913	0.018	0.288	...	0.000	0.000

TABLE VI. Normalized Feature Vectors of TI-LDA

Number	Dimension					
	1	2	3	...	2310	2311
1	0.662	0.350	0.000	...	0.000	0.000
	0.726	0.056	0.056	...	0.000	0.000
2	0.000	0.350	0.270	...	0.000	0.000
	0.042	0.042	0.043	...	0.000	0.000
3	0.662	0.000	0.270	...	0.000	0.000
	0.000	0.971	0.000	...	0.000	0.000
...	...	...	...	...	...	...
	...	...	...	...	...	...
1682	0.000	0.000	0.000	...	0.781	0.000
	0.913	0.019	0.420	...	0.000	0.000
1683	0.000	0.000	0.000	...	0.000	0.599
	0.913	0.018	0.288	...	0.000	0.000

For the three sets of vectors, this paper used the CFSFDP method for text clustering. According to the principle of CFSFDP, the relative distance values between the vectors was first calculated, and the results are shown in Table VII, Table VIII and Table IX. Then, we calculated the values of γ_i and drew them in descending order on a two-dimensional plane, as shown in Fig. 5, Fig. 6, and Fig. 7. Finally, we determined the number of clustering centers and classes based on the numerical variation diagram of γ_i. It's easy to see from Fig. 5, Fig. 6 and Fig. 7 that although the methods of feature extraction are different, the number of classes are the same.

TABLE VII. Relative Distance Values of TF-LDA

Number	Distance	Number	Distance
R_{11}	0.000	R_{598749}	2.796
R_{12}	1.883	R_{598750}	2.117
R_{13}	1.883	R_{598751}	2.362
R_{14}	1.883	R_{598752}	3.555
R_{15}	1.866	R_{598753}	3.709
...	...	...	...
R_{598743}	2.383	$R_{16831678}$	1.270
R_{598744}	3.593	$R_{16831679}$	1.247
R_{598746}	2.462	$R_{16831681}$	1.558
R_{598747}	2.283	$R_{16831682}$	1.983
R_{598748}	2.192	$R_{16831683}$	0.000

TABLE VIII. Relative Distance Values of LDA

Number	Distance	Number	Distance
R_{11}	0.000	R_{598749}	0.488
R_{12}	1.161	R_{598750}	0.393
R_{13}	1.161	R_{598751}	0.348
R_{14}	1.161	R_{598752}	1.176
R_{15}	1.067	R_{598753}	1.094
...	...	...	...
R_{598744}	1.128	$R_{16831679}$	0.956
R_{598745}	1.094	$R_{16831680}$	0.723
R_{598746}	0.553	$R_{16831681}$	0.219
R_{598747}	0.925	$R_{16831682}$	0.983
R_{598748}	2.209	$R_{16831683}$	0.000

TABLE IX. Relative Distance Values of TI-LDA

Number	Distance	Number	Distance
R_{11}	0.000	R_{598749}	4.087
R_{12}	2.637	R_{598750}	2.727
R_{13}	2.637	R_{598751}	3.598
R_{14}	2.637	R_{598752}	4.864
R_{15}	2.620	R_{598753}	3.709
...	...	...	...
R_{598744}	3.587	$R_{16831679}$	2.009
R_{598745}	3.698	$R_{16831680}$	1.738
R_{598746}	3.536	$R_{16831681}$	2.474
R_{598747}	2.292	$R_{16831682}$	2.968
R_{598748}	2.209	$R_{16831683}$	0.000

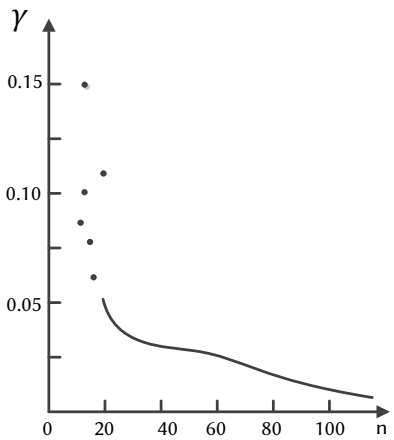

Fig. 5. γ_i Value Variation Diagram of TF-IDF.

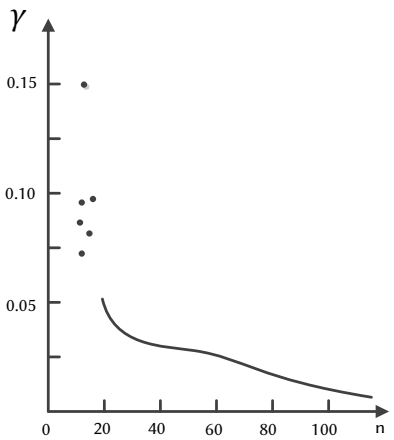

Fig. 6. γ_i Value Variation Diagram of LDA.

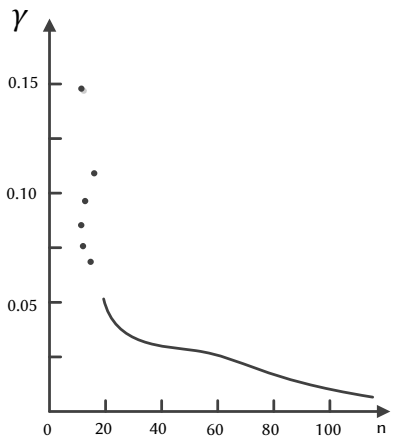

Fig. 7. γ_i Value Variation Diagram of TI-LDA.

To evaluate the clustering effect, this paper looks at two aspects, the intraclass compactness and the interclass separability, and uses the average intraclass compactness ($\overline{CP}$) and the average interclass separability ($\overline{SP}$) indicators to do the evaluation. The smaller the $\overline{CP}$'s value, the higher the compactness of the entire data set; the larger the $\overline{SP}$'s value, the higher the separability of the entire dataset is. To reflect the clustering effect on the entire dataset, this paper defines a comprehensive evaluation indicator $\overline{CS} = \overline{SP} / \overline{CP}$. It's easy to observe that the larger $\overline{CS}$'s value, the better the clustering comprehensive effect. The specific results are shown in Table X.

TABLE X. Comparison Table of Clustering Indicators

Indicator	Method		
	TF-IDF	LDA	TI-LDA
$\overline{CP}$	0.0014	0.0003	0.0002
$\overline{SP}$	1.6085	0.8628	1.0214
$\overline{CS}$	1148.9	2876.0	5107.0

In terms of intraclass compactness, based on Table X, because TI-LDA's $\overline{CP}$ value is the smallest, TI-LDA has the highest intraclass compactness and makes obvious improvements compared with TF-IDF. In terms of interclass separability, TF-IDF performs best because of the higher dimension of the feature vectors; it is followed in performance by TI-LDA and LDA. For overall performances, TI-LDA gains the highest marks and is far ahead of TF-IDF and LDA. Altogether, the TI-LDA method proposed in this paper, effectively improves the clustering effect.

B. Effectiveness Verification Experiment of Text-Driven Failure Diagnosis Model

Because the TI-LDA method proposed in this paper is better than TF-IDF and LDA based on Table X, the subsequent processing of the confirmatory experiments is based on the clustering results of TI-LDA. Through CFSFDP text clustering, the text data in this paper was divided into six completely different failure types. The first failure type is a transmitter failure; the second is a signal failure, which mainly is a signal problem of different monitors; the third is the failure of the aircraft's flight parameter indicators; the fourth is a generator failure, which is mainly caused by a generator overload and a signal failure; the fifth is engine failure; and the last is the failure caused by mechanical fatigue. The details are shown in Table XI.

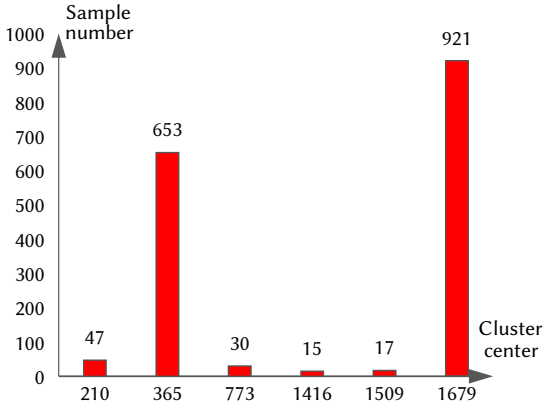

Fig. 8. Sample Number of Each Failure Type Under Original Data.

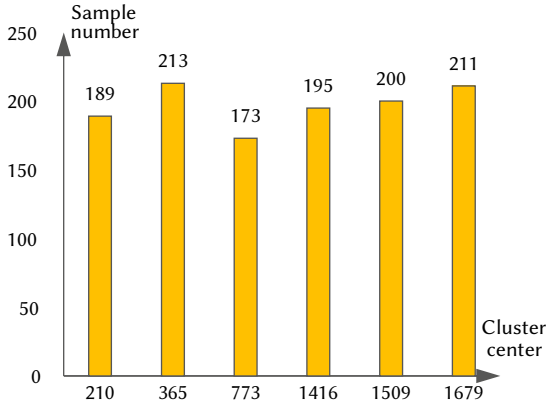

Fig. 9. Sample Number of Each Failure Type Under Oversampling Data.

In this paper, a number of samples of each failure type were calculated, as shown in Fig. 8. It can be seen from Fig. 8 that the text data presents an obvious class imbalance, with a large quantity gap between the major class and the minor class. In addition, the second type of failure and the sixth type of failure accounts for more than 90% of the failures, while the remaining four types accounts for a relatively small percent. In this paper, SMOTE is mainly used for data equalization to solve the class-unbalanced problem and to sample a number of each failure type after oversampling, as shown in Fig. 9.

TABLE XI. Clustering Results of TI-LDA

Class number	Clustering center	Text content of clustering center	Failure class
1	210	起飞后不能加高压，发射机故障 High voltage cannot be added after take-off and transmitter was damaged	Transmitter
2	365	电台不能发射，收不到信号，射频的K14a、b开关片地线断，23MZ的本振的L3电感器未并联上，16HZ频率调偏 The radio cannot transmit and the signal cannot be received. The ground wire of the switch pieces K14a and K14b of rf was broken. The L3 inductor of the 23MZ local vibration was not connected, and the frequency was deviated	Signal
3	773	校“I”状态记忆灯亮，地速指示极小（0021）且不动，低频分机故障 Status indicator light of “I” was on. The reading of the ground speed indicator was minimal (0021) and fixed. The low frequency extension broken down	Indicator
4	1416	1发起动发电机起动超负荷信号灯亮，减速器轴承漏光 The starting generator of the no. 1 engine was overloaded and the signal light was on. The reducer bearing was lightly leaking	Generator
5	1509	发动机停车后余油管大量漏油 Lots of oil leaked from the residual oil pipe after the engine stopped	Engine
6	1679	在“陆”位置加不上高压 High pressure cannot be applied at position “Land”	Mechanical fatigue

Comparing Fig. 8 and Fig. 9, the class imbalanced problem has been significantly improved and there is no significant difference in the number of samples between different classes after oversampling.

This paper used the original data and the oversampling data to train the SVM classifiers. By comparing the classification effect of the two classifiers, the effectiveness of the proposed text-driven failure diagnosis model was verified. Usually, the classification effect of classifier is mainly judged by the classification accuracy. However, for the text data with a class imbalance problem in this paper, the classification accuracy is not comprehensive. Therefore, this paper decided to use the confusion matrix from *FNR* (False Negative Rate), *FPR* (False Positive Rate), *Acc* (Accuracy), *Recall* and F_1 to evaluate the classification effect. The specific results are shown in Table XII.

TABLE XII. Comparison TABLE of Evaluation Indicators of SVM Classifier

Indicator	Original data	Oversampling data
FNR	0.149	0.021
FPR	0.015	0.008
Acc	0.982	0.989
Recall	0.851	0.979
F_1	0.721	0.968

From Table XII, it can be shown that the classifier trained by the original data is very close to the classifier trained by the oversampling data in *Acc*, and both have high accuracy, which further reflects the validity and feasibility of the proposed failure diagnosis model from a data perspective. However, for other classification indicators, the failure diagnosis model trained by the oversampling data has obvious improvements in *Recall* and F_1, which reflects that there exist false high accuracies of the classifier trained by the original data. According to the characteristics of the data in this paper, the major class usually has many more samples than the minor class. Therefore, classifiers tend to group minor-class samples into the major class, which often results in high false alarm rates of the major class and high missing alarm rates of the minor class. This can also be seen from the *FNR* and *FPR*. The *FNR* and *FPR* of the classifier trained by the oversampling data are much smaller, so there are fewer mistakes in the classification task, which effectively avoids the high false alarm rates.

VI. Conclusion and Discussion

In a variation from the traditional methods of failure diagnoses based on structured data, this paper proposes a text-driven failure diagnosis model by using NLP technology, which fills in a gap for research on failure diagnoses based on unstructured data, especially text data.

To resolve the shortcomings of traditional TF-IDF and LDA text feature extraction methods, this paper proposes TI-LDA, a new text feature extraction method, based on serial feature fusion, and uses the CFSFDP clustering method to verify the effectiveness of TI-LDA. The final experimental results show that the feature vectors extracted by TI-LDA can effectively improve intraclass compactness and interclass separability, compared with the methods using TF-IDF and LDA alone.

In this paper, the TI-LDA method is applied to the field of failure diagnosis, and a text-driven failure diagnosis model based on TI-LDA is created by combining the machine learning methods such as CFSFDP clustering, SMOTE oversampling and SVM classification. This failure diagnosis model can classify and automatically judge which failure mode the failure described in the text belongs to once a failure-description text is entered. Through an effectiveness verification experiment, it was found that the failure diagnosis model proposed in this paper has a high accuracy, and effectively solves the problem of high false accuracies and high false alarm rate caused by the class imbalance problem.

It's worth mentioning that this text-driven failure diagnosis model is unsupervised, which means it does not need any label data, so it has better portability and lower labor costs. This model has very broad application prospects, especially in the failure diagnosis field for large and complex equipment such as for aircraft, high-speed rail, and even for the medical field.

However, the model can still be improved. For example, the TI-LDA text feature extraction method fuses according to a ratio of 1:1, but this ratio may be not the best fusion ratio. So, our next work will study how to find the best fusion ratio to obtain the best fusion effect. In addition, in order to simplify the study, this paper does not consider the interference of abnormal or unreal text data. Future work will focus on ways to identify abnormal text data so as improve the accuracy of the model.

Acknowledgment

This work is supported by the National Natural Science Foundation of China (Grant No.71971013 and 71871003) and the Fundamental Research Funds for the Central Universities (YWF-20-BJ-J-943). The study is also sponsored by the Aviation Science Foundation of China(2017ZG51081), the Civil Aircraft Science Research Fund (MJ-2017-J-92) and the Graduate Student Education and Development Foundation of Beihang University.

References

[1] Cambria, E. & White, B., "Jumping NLP Curves: A Review of Natural Language Processing Research [Review Article]", Computational Intelligence Magazine IEEE, vol. 9, no. 2, pp. 48-57, 2014.

[2] Tom, Y., Devamanyu, H., Soujanya, P. & Erik, C., "Recent Trends in Deep Learning Based Natural Language Processing [Review Article]", IEEE Computational Intelligence Magazine, vol. 13, no. 3, pp. 55-75, 2018.

[3] Salton, G., Wong, A. & Yang, C., "A vector space model for automatic indexing", Communications of the ACM, vol. 18, no. 11, pp. 613-620, 1975.

[4] Liang, H., Sun, X., Sun, Y. & Gao, Y., "Text feature extraction based on deep learning: a review", EURASIP journal on wireless communications and networking, vol. 2017, no. 1, pp. 1-12, 2017.

[5] Sparck Jones, K., "A statistical interpretation of term specificity and its application in retrieval", Journal of documentation, vol. 28, no. 1, pp. 11-21, 1972.

[6] Blei, D. M., Ng, A. Y. & Jordan, M. I., "Latent dirichlet allocation", Journal of machine Learning research, vol. 3, no. Jan, pp. 993-1022, 2003.

[7] Wang, F., Xu, T., Tang, T., Zhou, M. & Wang, H., "Bilevel feature extraction-based text mining for fault diagnosis of railway systems", IEEE transactions on intelligent transportation systems, vol. 18, no. 1, pp. 49-58, 2016.

[8] Rajpathak, D. G. & Singh, S., "An Ontology-Based Text Mining Method to Develop D-Matrix From Unstructured Text", IEEE Transactions on Systems Man & Cybernetics Systems, vol. 44, no. 7, pp. 966-977, 2014.

[9] Rodrigues, R. E. R. S., Balestrassi, P. P., Paiva, A. P., Garcia-Diaz, A. & Pontes, F. J., "Aircraft interior failure pattern recognition utilizing text mining and neural networks", Journal of Intelligent Information Systems, vol. 38, no. 3, pp. 741-766, 2012.

[10] Harris, Z. S., "Distributional structure", Word, vol. 10, no. 2-3, pp. 146-162, 1954.

[11] Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K. & Harshman, R., "Indexing by latent semantic analysis", Journal of the American society for information science, vol. 41, no. 6, pp. 391-407, 1990.

[12] Roweis, S. T. & Saul, L. K., "Nonlinear dimensionality reduction by locally linear embedding", science, vol. 290, no. 5500, pp. 2323-2326, 2000.

[13] Bengio, Y., Ducharme, R. E. J., Vincent, P. & Jauvin, C., "A neural probabilistic language model", Journal of machine learning research, vol. 3, no. Feb, pp. 1137-1155, 2003.

[14] Ming, T., Lei, Z., Xianchun, Z. & Others, "Document vector representation based on Word2Vec", Computer Science, vol. 43, no. 6, pp. 214-217, 2016.
[15] Adali, T. U. L., Levin-Schwartz, Y. & Calhoun, V. D., "Multimodal data fusion using source separation: Application to medical imaging", Proceedings of the IEEE, vol. 103, no. 9, pp. 1494-1506, 2015.
[16] Jing, L., Wang, T., Zhao, M. & Wang, P., "An adaptive multi-sensor data fusion method based on deep convolutional neural networks for fault diagnosis of planetary gearbox", Sensors, vol. 17, no. 2, pp. 414, 2017.
[17] Wang, S., Deng, Z. & Yin, G., "An accurate GPS-IMU/DR data fusion method for driverless car based on a set of predictive models and grid constraints", Sensors, vol. 16, no. 3, pp. 280, 2016.
[18] Khaleghi, B., Khamis, A., Karray, F. O. & Razavi, S. N., "Multisensor data fusion: A review of the state-of-the-art", Information fusion, vol. 14, no. 1, pp. 28-44, 2013.
[19] Yang, J., Yang, J., Zhang, D. & Lu, J., "Feature fusion: parallel strategy vs. serial strategy", Pattern recognition, vol. 36, no. 6, pp. 1369-1381, 2003.
[20] Liu, C. & Wechsler, H., "A shape-and texture-based enhanced Fisher classifier for face recognition", IEEE transactions on image processing, vol. 10, no. 4, pp. 598-608, 2001.
[21] Sun, Q., Zeng, S., Liu, Y., Heng, P. & Xia, D., "A new method of feature fusion and its application in image recognition", Pattern Recognition, vol. 38, no. 12, pp. 2437-2448, 2005.
[22] Geman, S. & Geman, D., "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images", IEEE Transactions on pattern analysis and machine intelligence, no. 6, pp. 721-741, 1984.
[23] Visser, E., Nijhuis, E. H., Buitelaar, J. K. & Zwiers, M. P., "Partition-based mass clustering of tractography streamlines", NeuroImage, vol. 54, no. 1, pp. 303-312, 2011.
[24] Bouguettaya, A., Yu, Q., Liu, X., Zhou, X. & Song, A., "Efficient agglomerative hierarchical clustering", Expert Systems with Applications, vol. 42, no. 5, pp. 2785-2797, 2015.
[25] Tang, X. & Zhu, P., "Hierarchical clustering problems and analysis of fuzzy proximity relation on granular space", IEEE Transactions on Fuzzy Systems, vol. 21, no. 5, pp. 814-824, 2012.
[26] Lu, J. & Zhu, Q., "An effective algorithm based on density clustering framework", Ieee Access, vol. 5,4991-5000, 2017.
[27] Tu, L. & Chen, Y., "Stream data clustering based on grid density and attraction", ACM Transactions on Knowledge Discovery from Data (TKDD), vol. 3, no. 3, pp. 1-27, 2009.
[28] Rodriguez, A. & Laio, A., "Clustering by fast search and find of density peaks", Science, vol. 344, no. 6191, pp. 1492-1496, 2014.
[29] Jain, A. K., "Data clustering: 50 years beyond K-means", Pattern recognition letters, vol. 31, no. 8, pp. 651-666, 2010.
[30] Japkowicz, N. & Stephen, S., "The class imbalance problem: A systematic study", Intelligent data analysis, vol. 6, no. 5, pp. 429-449, 2002.
[31] Chawla, N. V., Bowyer, K. W., Hall, L. O. & Kegelmeyer, W. P., "SMOTE: synthetic minority over-sampling technique", Journal of artificial intelligence research, vol. 16,321-357, 2002.
[32] Cortes, C. & Vapnik, V., "Support-vector networks", Machine learning, vol. 20, no. 3, pp. 273-297, 1995.

Shenghan ZHOU

Shenghan ZHOU, Doctor of management science, Associate Professor, graduated from the Beihang University in 2009, working in Beihang University. His research interests include system safety, data mining, and risk management.

Bang CHEN

Bang CHEN was born in Xinyang, Henan, China in 1995. He received the B.S. degree in safety engineering from Northeastern University Shenyang, Liaoning, China, in 2019. He is currently pursuing the M.S. degree in safety science and engineering at Beihang University, Beijing China. His research areas are data mining and nature language processing.

Yue ZHANG

Yue ZHANG was born in Shanxi Province, China. 1997. She received the B.S. degree in engineering management from Nanjing Agricultural University, Nanjing, China, in 2019. She is currently pursuing the M.S. degree in safety science and engineering at Beihang University, Beijing, China. Her research areas are data mining, factory layout and path optimization.

HouXiang LIU

HouXiang LIU was born in Anhui Province, China. 1997. He received the B.S. degree in safety engineering from Northeastern University, Shenyang, China in 2019. He is currently pursuing the M.S. degree in industrial engineering at Beihang University, Beijing, China. His research areas are data analysis and traffic optimization.

Yiyong XIAO

Yiyong XIAO, Doctor of System Engineering, Associate Professor, graduated from the Beihang University in 2003. Working in Beihang University. His research interests include economic affordability, data mining, network security optimization and algorithm research.

Xing PAN

Xing PAN graduated the doctor degree in management science from the Beihang University, in 2005, where he is currently a Senior Engineer. His research interests include systems engineering, data mining, and risk assessment.