Hindawi
Complexity
Volume 2024, Article ID 9827214, 1 page
https://doi.org/10.1155/2024/9827214

WILEY | Hindawi

Retraction

Retracted: Application Research of Intelligent Classification Technology in Enterprise Data Classification and Gradation System

Complexity

Received 23 January 2024; Accepted 23 January 2024; Published 24 January 2024

This article has been retracted by Hindawi following an investigation undertaken by the publisher [1]. This investigation has uncovered evidence of one or more of the following indicators of systematic manipulation of the publication process:

(1) Discrepancies in scope

(2) Discrepancies in the description of the research reported

(3) Discrepancies between the availability of data and the research described

(4) Inappropriate citations

(5) Incoherent, meaningless and/or irrelevant content included in the article

(6) Manipulated or compromised peer review

The presence of these indicators undermines our confidence in the integrity of the article's content and we cannot, therefore, vouch for its reliability. Please note that this notice is intended solely to alert readers that the content of this article is unreliable. We have not investigated whether authors were aware of or involved in the systematic manipulation of the publication process.

Wiley and Hindawi regrets that the usual quality checks did not identify these issues before publication and have since put additional measures in place to safeguard research integrity.

We wish to credit our own Research Integrity and Research Publishing teams and anonymous and named external researchers and research integrity experts for contributing to this investigation.

The corresponding author, as the representative of all authors, has been given the opportunity to register their agreement or disagreement to this retraction. We have kept a record of any response received.

References

[1] L. Yu, C. Wang, H. Chang, S. Shen, F. Hou, and Y. Li, "Application Research of Intelligent Classification Technology in Enterprise Data Classification and Gradation System," *Complexity*, vol. 2020, Article ID 6695484, 9 pages, 2020.

Hindawi
Complexity
Volume 2020, Article ID 6695484, 9 pages
https://doi.org/10.1155/2020/6695484

Research Article

Application Research of Intelligent Classification Technology in Enterprise Data Classification and Gradation System

Lina Yu,[1] **Chunwei Wang**,[1,2] **Huixian Chang**,[1] **Sheng Shen**,[1] **Fang Hou**,[1] **and Yingwei Li**[1]

[1]*School of Information Science and Engineering, Yanshan University, Qinhuangdao 066004, China*
[2]*Beijing Branch, Daqing Oilfield Information Technology Company, Beijing 100043, China*

Correspondence should be addressed to Yingwei Li; lyw@ysu.edu.cn

Received 22 October 2020; Revised 18 November 2020; Accepted 25 November 2020; Published 7 December 2020

Academic Editor: Zhihan Lv

Classification and gradation system adopts different security protection schemes for different types of data by implementing classification and gradation management of data, which is an important pretechnical means for data security protection and prevention of data leakage. This paper introduces artificial intelligence classification, machine learning, and other means to learn and train enterprise documents according to the characteristics of enterprise sensitive data. The generated training model can intelligently identify and classify file streams, improving work efficiency and accuracy of classification and gradation. At the same time, the differences, advantages, and disadvantages of K-NN (K-Nearest Neighbors), DT (Decision Tree), and LinearSVC algorithms are compared. The experimental data shows that LinearSVC algorithm is applicable to high-dimensional data, with discrete, sparse data features and large number of features, which is more suitable for classification of sensitive data of enterprises.

1. Introduction

With the advent of Internet plus era, the status of data as a basic strategic resource has become increasingly prominent [1–3]. As network security threats become increasingly prominent, data face many security risks during storage, processing, and transmission. The challenges of management and monitoring which accompany it are increasingly severe [4, 5], so the establishment of a data classification and gradation system environment is the basic premise for the rapid development of enterprises [6, 7]. According to the results of data classification and gradation and the company's policy requirements, efficiently matching the collated data with security strategies is an important means for enterprises to improve their core competitiveness. For large enterprises, data security faces many problems and challenges. According to the international standards of information security [8], the importance of different data is different, and high-value data requires stricter protection mechanisms. Therefore, data is used as the security protection target, and the intricate enterprise data assets are divided into various categories and multiple levels according to the classification and gradation method. According to the type and value of data, different protection strategies are formulated [9, 10], and the continuous strengthening and improvement of sensitive data security management have become more prominent and important.

At present, the state has promulgated laws and regulations such as the "Network Security Law of the People's Republic of China" [11], clearly stating that "network operators should comply with the requirements of the network security level protection system, perform the following security protection obligations to protect the network from interference, destruction, or unauthorized access, prevent network data from being leaked, stolen, or tampered," and "take measures such as data classification, important data backup, and encryption". The data combining work has the following problems. First, for the sensitive data of enterprises, due to the lack of strict management of the computer network by the managers and their lack of awareness of network security, it is easy for the computer to leak information, which causes a series of subsequent security risks to damage the computer network [12]. Second, most of the understanding of policies, regulations, and standards also

relies on manual combining. There is a possibility that the understanding and interpretation of policies, regulations, and standards may be artificially expanded or reduced. Third, at the same time, with regard to the characteristics of large data business types and large data volumes of large enterprises [13, 14], business personnel cannot quickly identify which sensitive data is based on standards, and the level of confidentiality corresponding to sensitive data is not easy to define. Therefore, establishing a classification and gradation protection catalogue for the corresponding trade secrets of large enterprises, establishing a more detailed classification and gradation process specification and implementation guide according to the business data, improving the efficiency and accuracy of classification and gradation management, providing reference for confidentiality of documents for business personnel, and forming compliance management measures are imperative [15, 16].

By integrating artificial intelligence classification technology [17], this paper carries out automatic classification and gradation of documents. This paper compares three classification algorithms and shows the feasibility and effect of LinearSVC algorithm in the classification and gradation system, so as to improve the accuracy of data classification and gradation [18].

2. Classification and Gradation System Architecture and Deployment

The classification and gradation management of enterprise information can enable various sensitive data information of the enterprise to be grasped in a timely, efficient, and accurate manner [19]. It is an important pretechnical means for large-scale enterprises to protect data security and prevent data leakage. According to the data privacy standards and management systems of enterprises, classification and gradation management of sensitive data of enterprises is more convenient for the formulation of data security policies and the protection of sensitive data.

Figure 1 shows the deployment of classification and gradation. As the core part of classification and gradation, AI server can predict the classification and gradation of enterprise documents through artificial intelligence classification and machine learning and provide users with a higher level of intelligent application services; the role of the load-balancing server is to distribute the user to the web server group; the file server and the database, respectively, provide storage for the file and relational data for the classification and gradation system; and the report server provides a comprehensive display for the classification and gradation information.

The DLP system continuously captures and analyzes the traffic on the network by placing a monitor at the outlet of the enterprise's external network and detects sensitive data and important traffic elements [20] through protocols such as SMTP, FTP, and HTTP to prevent the transfer of sensitive data to the outside. It can be seen from Figure 2 that the classification and gradation system integrates with the DLP (data leakage prevention) system to obtain network or terminal data leakage event information and displays related information for organizations with data permissions on the basis of strict authority control design. The user data protection requirements can be collected from the information or security management department to provide a basis for the data security policy. At the same time, the DLP system interacts with the classification and gradation system to compare the fingerprint of the outgoing file with the fingerprint database of the classification and gradation system to predict which category the file belongs to.

3. Materials and Methods

Data intelligent classification mainly uses intelligent classification technology to form different categories of data to be classified [21]. As shown in Figure 3, the AI intelligent classification function is mainly divided into two modules: AI training and AI classification. The AI training module is processed by a stand-alone server. Through the full amount of learning from classification and gradation information source uploaded by users, generating the system model, and uploading it to AI classification module, AI classification will classify documents according to the model. In the AI classification module, when the user enters the classification and gradation information, the platform intelligently classifies and gives the user a classification prompt to provide a reference for the user.

3.1. Data Preprocessing. To be able to calculate the accuracy of the classification algorithm, before the model training, the system needs to automatically classify and annotate the data sample files and then classify the inaccurate files automatically into the correct classification through manual proofreading. The manual proofreading step is very important and has a great impact on the final accuracy. For the calibration document after proofreading, if obtaining higher quality corpus data is needed during model training, the data should be preprocessed in advance [22]. The data preprocessing flow chart is shown in Figure 4.

The original text is first cleaned. This is the last procedure to find and correct identifiable errors in the data file. It reexamines and verifies the data, removing duplicate information, correcting existing errors, and providing data consistency. After that, the word segmentation process is performed, the Chinese characters are divided into individual words, and the continuous word sequences are recombined into word sequences according to certain specifications. Finally, the stop words are removed. The words that do not contribute to the text features are roughly divided into two categories. One kind of stop words is characterized by a wide range of applications and can be found everywhere in various documents; for example, the word "company" appears in almost every document, and the characteristics of the document cannot be reflected for such words. Another kind of stop words includes a modal particle, an adverb, a preposition, and a conjunction, which usually has no clear meaning. These words will not have specific meaning until they are put into a complete sentence, such as the common "the," "in," and the like. After the data is

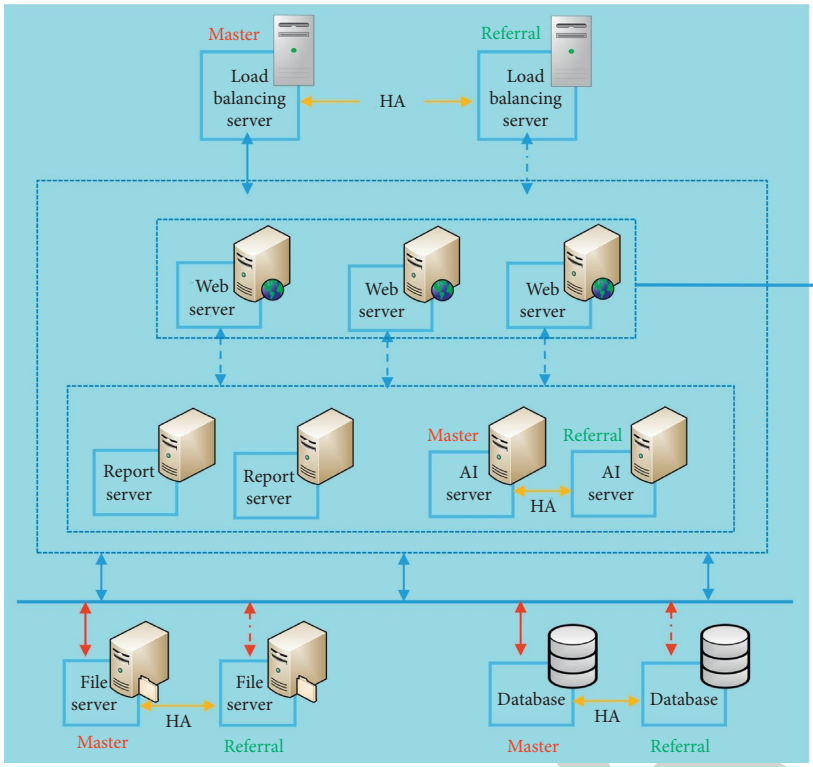

Figure 1: Classification and gradation system architecture diagram.

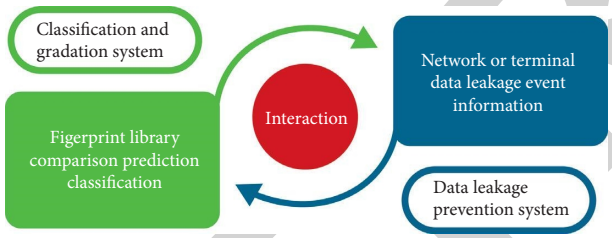

Figure 2: Interaction diagram of classification and gradation system and DLP system.

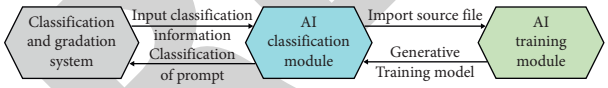

Figure 3: AI intelligent classification module diagram.

preprocessed, the original text is expected to have higher quality prediction data for the next model training.

3.2. Model Training. Dividing all training data into two parts, one for the training model is called the train set, and the other for the accuracy test of the model is called the test set. Then the TF-IDF (term frequency-inverse document frequency) [23] calculation on the two data sets is carried out. The calculation process is as follows: for the word t_i in each parsed.txt file d_j, the term frequency (TF) can be expressed as in the following equation:

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

where $n_{i,j}$ is the number of occurrences of the term t_i in the document d_j and the denominator is the sum of the occurrences of all the terms in the document d_j. The main idea of IDF (inverse document frequency) is that if there are fewer documents containing the word t_i, the IDF value is larger, indicating that the word t_i has a good class discrimination ability at the level of the entire document set. IDF is expressed as follows:

$$\mathrm{idf}_i = \log \frac{|D|}{\left|\left\{j : t_i \in d_j\right\}\right|}, \tag{2}$$

where $|D|$ is the total number of all documents in the corpus and the denominator is the number of all documents containing the term t_i. The TF-IDF weight is actually the product of two parameters. That is, $\mathrm{tf}_{i,j} \times \mathrm{idf}_i$.

TF-IDF is a commonly used weighting technique for information retrieval and data mining. It is a statistical method used to evaluate the importance of a word to a document set or one of the documents in a corpus [24]. The importance of a word increases in proportion to the number of times it appears in the document, but at the same time it decreases in inverse proportion to the frequency of its appearance in the corpus. If in a specific type of test set document the TF-IDF value of the word t_i is high but that in other categories is very low or even 0, indicating that the word is of greater importance to this type of document and has a strong ability to classify this type of document, it can be

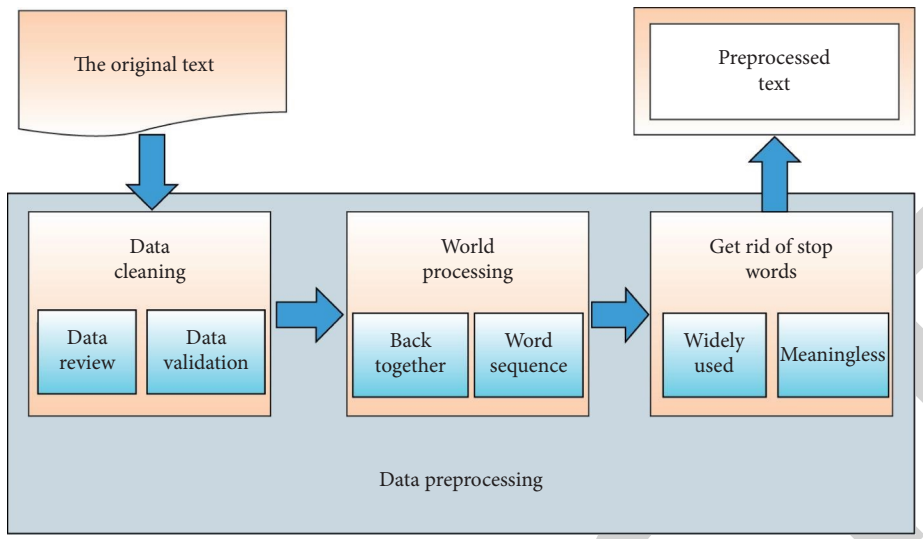

Figure 4: Data preprocessing flow chart.

regarded as the characteristic word of this kind of file. The specific model training process is as follows: TF-IDF calculation is performed on the words in the file data in each category of the training set; then the words with the strongest classification ability are selected as the feature words of the type of file. So, the classifier training is performed using these features. In the end, a classification model is generated. After each training of the model, the test data set needs to be used to verify the accuracy of the model. When the accuracy reaches the requirement, the model can be saved. If the higher accuracy is not achieved, it is necessary to classify the annotation data to confirm whether there is a data classification error and then adjust the parameters of the model, evaluate the model, and iterate the process until a higher accuracy is obtained. The model training process is shown in Figure 5.

3.3. Learning Function of Classification Model. In view of the numerous and intricate features of the documents in the enterprise, with the change of the enterprise system and the continuous increase of documents, it is sometimes necessary to make changes to the classification standards and adjust the data file categories. The AI intelligent classification model learning framework is shown in Figure 6. The framework consists of model learning services, file parsing services, and model training. With the corpus and new standard of classification reuploaded, the file parsing service is called to convert the data file into a txt file, and the corpus is initially automatically classified based on the new classification criteria. After manual proofreading, the model training module is invoked for training. The result is repeated iterations to achieve the learning function of the model.

4. Results and Discussion

4.1. Several Common Classification Algorithms. Logistic Regression (LR) [25], a generalized linear regression analysis model, is based on the Sigmoid function to deal with large-scale data by giving the probability that the sample belongs to each category. Logistic regression algorithms are often used to solve the two classification problems; Naive Bayes (NB) [26], a method for calculating posterior probabilities from prior probabilities, requires a hypothetical premise. In the actual data classification analysis process, this premise assumption is often too idealistic, and it is not established in the actual situation. So the above two algorithms are not suitable for use in the enterprise data multiclassification system.

Decision Tree (DT) [27] creates a tree node by calculating the information gain of each attribute and selecting the attribute with the highest information gain as the test attribute of the given data set, and marks with this attribute, and then creates a separate branch for each value of the attribute and divides the sample accordingly; K-Nearest Neighbors (K-NN) [28] classification, an analogy-based learning method, works by storing all training samples in an N-dimensional model space, calculating sample files by calculating the K training samples closest to a given unknown sample using the Euclidean distance formula; LinearSVC [29] is one of the SVM (Support Vector Machine) classification algorithms. By using kernel function techniques, linearly inseparable features are mapped into high-dimensional space, so that features can be divided in high-dimensional space. Based on the limited sample information, the complexity of the model (the learning accuracy of a particular training sample) and the learning ability (the ability to identify any sample without error) maximize the maximum separation between the separate categories to achieve good classification predictions for the sample file.

4.2. Experimental Scheme and Results. Intelligent classification and gradation module is essentially a module of text classification. Text classification refers to the automatic classification process of the input text according to a certain categorization system by the computer through algorithms. The algorithm of this classification and gradation module is implemented by a more mature machine learning algorithm.

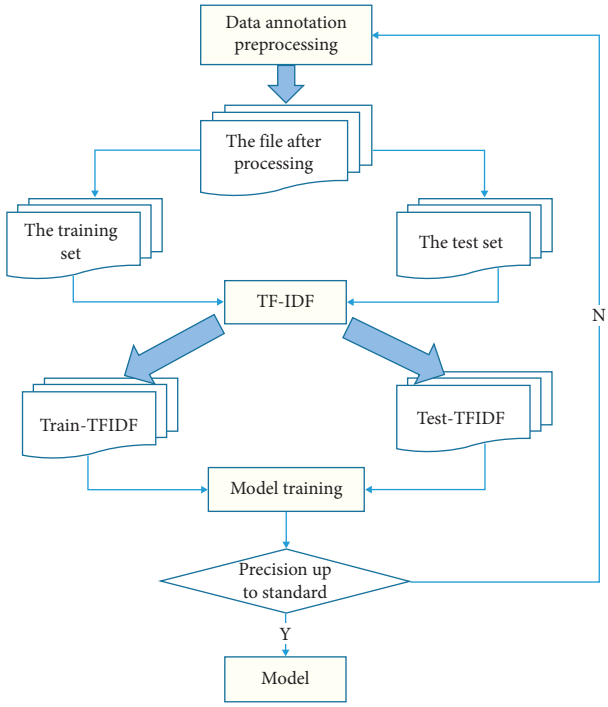

Figure 5: Model training flow chart.

In machine learning, there are many algorithms that can be used for text classification. After comparing the advantages and disadvantages of different algorithms, combined with the sparse and discrete features of the enterprise text data, three algorithms (Decision Tree, K-NN, and LinearSVC) that are suitable for enterprise text classification are selected for experiment.

For the experimental work of exploring the data volume of the model training on the accuracy and modeling time of the three classification algorithms, according to most enterprise systems and organizational components, the enterprise data can generally be roughly divided into 12 categories including personnel, auditing, legal affairs, material procurement, production and management, technology management, discipline inspection and supervision, maintenance of letters and visits, comprehensive office, planning, finance, international cooperation, and policy research. For the discrete, sparse, and feature-rich characteristics of the enterprise data, if there are too many training samples, some abnormal feature values will appear inside for the discrete and sparse enterprise data, which will affect the accuracy of the model. At the same time, the amount of training data should not be too small; otherwise, it will also affect the accuracy. In view of the fact that the number of training samples would affect the accuracy of the classification model, this experiment fixed the size of the sample categories into 12 to explore the changes in the accuracy and modeling time of the three classification algorithms when the amount of training data for each classification file is 40, 60, 80, 100, 120, 140, and 160. It is worth noting that, in order to avoid the influence of other factors on the experiments, all experiments were conducted on the same hardware configuration server. The results are shown in Figure 7.

Due to the characteristics that the enterprise data are discrete, sparse, and numerous, the more categories are classified, the higher the degree of feature coincidence among each category is, which makes the accuracy of the model more easily affected. In general, the enterprise documents are roughly divided into 12 categories: company personnel, finance, and so on. According to the different nature and the different system of the enterprise, the data category species will slightly change.

When exploring the influence of the number of training samples on the model accuracy and modeling time, it can be seen from Figure 7 that the modeling time increases with the increase of the number of training samples, while the accuracy of the model built by the three algorithms peaks within the range of local sample size (75–100). Therefore, as for the experimental work to explore the influence of the number of enterprise data classification types on the accuracy and modeling time of the three classification algorithms, in order to facilitate the calculation of experimental parameters, the fixed training sample size for this experiment is 100 files. Therefore, this experiment explores the changes in the accuracy and modeling time of the three classification algorithms when the number of classification types is 8, 10, 12, 14, 16, 18, and 20 in sequence for servers with the same hardware configuration. The variations of the three algorithms are displayed in Figure 8.

4.3. Analysis of Experimental Results. It can be seen from the data and the line graph obtained from the two experiments that, for different numbers of training samples, the model training time of each algorithm is almost the same for the three algorithms. But the training time of the Decision Tree algorithm model is generally over 1s, while the other two algorithms K-NN and LinearSVC have slightly faster speed, with the over 1s training time when the training sample size is more than 100. However, in the view of the accuracy, LinearSVC classification algorithm is superior to the other two kinds of algorithm; its accuracy can reach 95% or so. The LinearSVC classification algorithm has the highest accuracy when the number of training samples is about 100 files. For the experiments for different types of enterprise data classification, the training times of all algorithms are still similar, but the accuracy of the LinearSVC classification algorithm is still the highest, which can reach about 95%. According to the results of the two experiments, it can be seen that, when considering the time used for modeling, the three classification algorithms are not much different, but the LinearSVC classification algorithm still has a far better accuracy than the other two algorithms in the two experiments and therefore is most suitable for application in the enterprise data classification and gradation system.

5. Application Verification and Summary

5.1. Analysis of Experimental Results. Taking a petroleum enterprise as an example, the AI intelligent classification composed of the LinearSVC algorithm is applied in the classification and gradation system of the enterprise. Then,

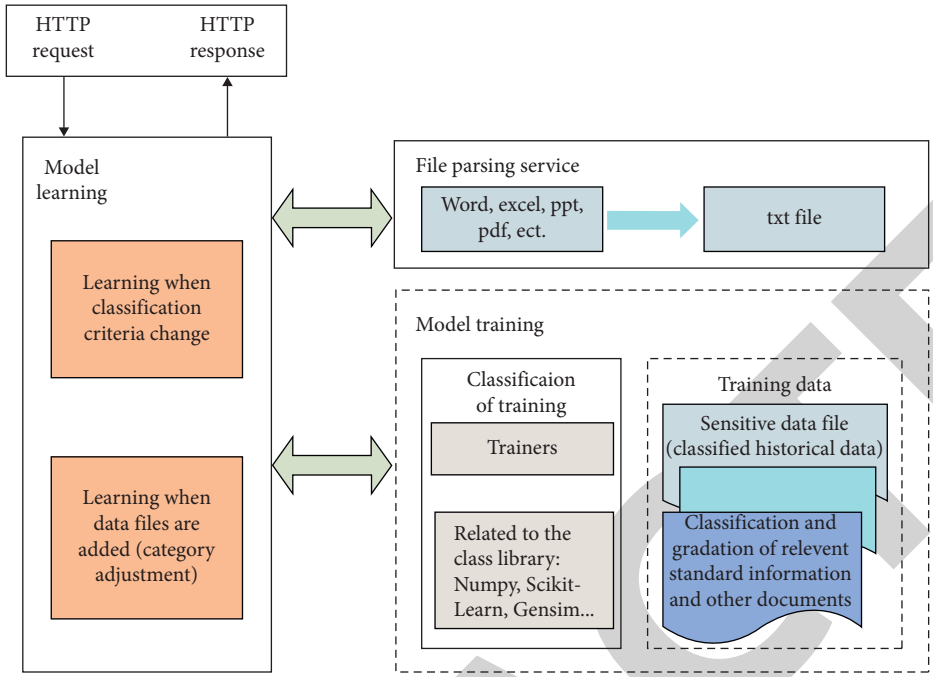

Figure 6: AI intelligence classification model learning frame diagram.

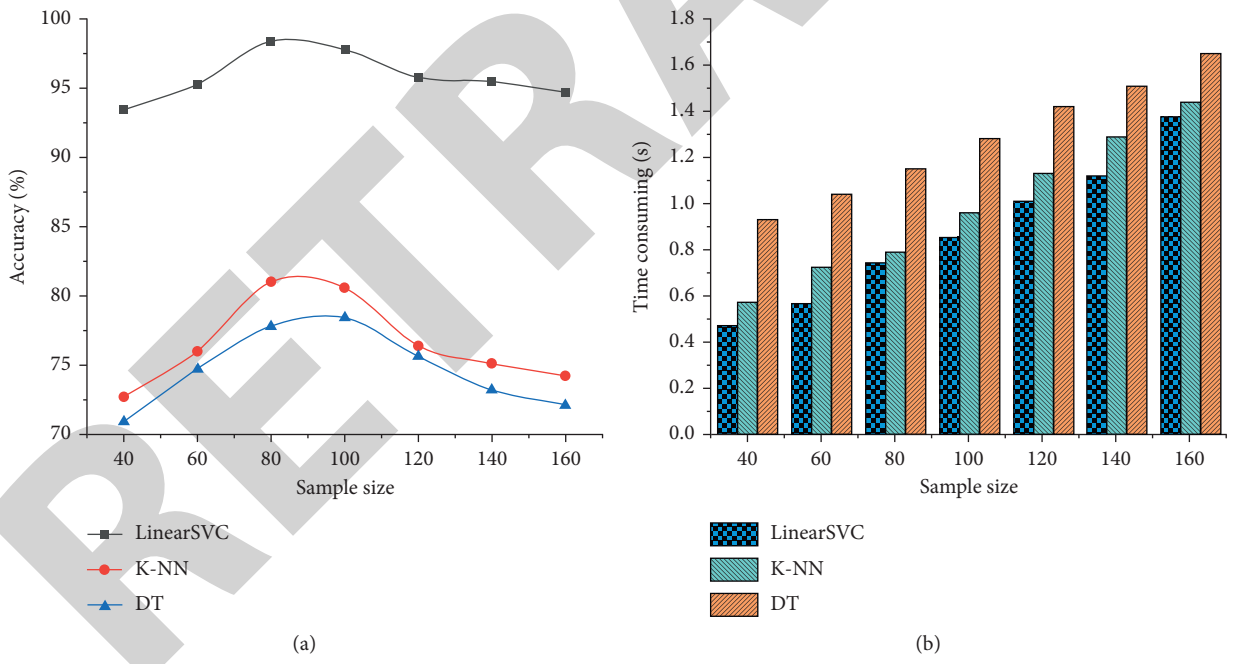

Figure 7: The relationship between different number of training samples and accuracy and time. (a) Accuracy line graph. (b) Modeling time diagram.

according to the nature and system of the oil company, the company roughly divides the data into 19 categories: personnel, auditing, legal affairs, and so on. At the same time, we found 19 types of documents from within the oil company, and the number of each type of documents is about 100. The various documents are divided into two parts: training set and test set. The number of the two text sets can be flexibly set, usually set as the ratio of training set and test set is 4 : 1, but the slight increase or decrease of this ratio will not affect the accuracy of model training results.

The numbers of data files for the training and test sets of each classification category are shown in Tables 1 and 2.

The supervised model training is carried out in the enterprise classification and gradation system. When the

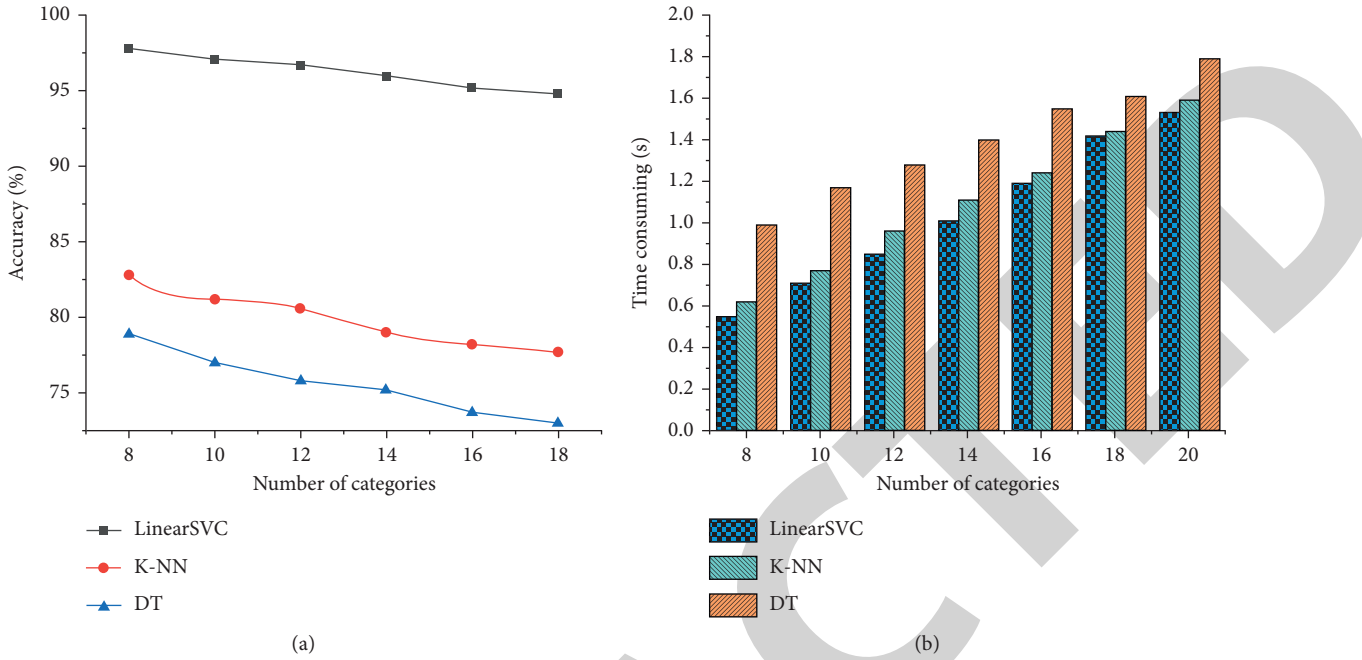

Figure 8: The relationship between training samples of different classification categories and accuracy and time. (a) Accuracy line graph. (b) Modeling time diagram.

Table 1: Number of files in each training set.

Category	Quantity
Professional branch	71
Personnel	64
Safety and environmental protection	60
Global cooperation	65
Information management	62
Audit	71
Reform	50
Policy research	61
Legal affairs	52
Material procurement	83
Production management	68
Technology management	99
Discipline inspection and supervision	53
Letter of visit	51
General office	49
Planning	65
Finance	76
Quality and standards	56
Capital operation	61
Total	1217

Table 2: Number of files in each test set.

Category	Quantity
Professional branch	37
Personnel	29
Safety and environmental protection	21
Global cooperation	31
Information management	21
Audit	25
Reform	19
Policy research	23
Legal affairs	27
Material procurement	37
Production management	31
Technology management	41
Discipline inspection and supervision	33
Letter of visit	37
General office	22
Planning	31
Finance	23
Quality and standards	24
Capital operation	26
Total	538

accuracy reaches the required level, the model is generated. Table 3 shows the accuracy and time taken for the training of the enterprise classification and gradation system model.

It can be seen from Table 3 that when the training accuracy reaches 100%, the time spent in the model training is less than 1s, and when tested, the accuracy of the model can reach 94%, exceeding 90%. Therefore, the trained model shows the advantages of short training time and high accuracy.

5.2. Application Verification. In order to verify the accuracy of the classification model of the system for the classification of sensitive data documents of enterprises, 3000 data files are randomly selected as test files in a petroleum enterprise. The test files were randomly divided into three groups, and the system was tested three times in succession and verified by three indicators: (3), (4), and (5).

Table 3: Model training results.

Category	Accuracy (%)	Model training time (s)
Training set	100	0.85
Test set	94.8	×

Table 4: Accuracy indicators of classification model.

Test number	Precision (%)	Recall (%)	F1 (%)
Test 1	94.9	95.3	95.1
Test 2	95.2	95.4	95.3
Test 3	94.3	95.0	94.6

$$\text{precision} = \frac{\text{Num}_{\text{correct}}}{\text{Num}_{\text{total}}}, \tag{3}$$

$$\text{recall} = \frac{\text{Num}_{\text{correct}}}{\text{Num}_{\text{actual}}}, \tag{4}$$

$$F1 = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}. \tag{5}$$

In the previous equations, $\text{Num}_{\text{correct}}$ is the correct number of documents in the various types of data identified by the classification model, $\text{Num}_{\text{total}}$ is the number of documents predicted by the model, and $\text{Num}_{\text{actual}}$ is the total number of actual samples.

Table 4 clearly reveals that, in the three random testing experiments, the generated model shows great classification effect: the recall rate reaches 95% and above, and the classification accuracies are all higher than 94%. The results are consistent with the conclusions drawn in the test section, indicating that the model has good stability and can be well applied to the classification of enterprise sensitive data documents.

Regarding the part of the misclassified data that has not been verified, although the data is not classified into the correct category, it will not affect the detection of sensitive files in actual scenario. The reason is as follows: With the increase in the classification categories of sensitive documents, the difference in classification standards between the categories will gradually decrease. Therefore, in this case, it is easy to cause the misclassification of data. But, in fact, even if the data is misclassified into other categories, due to the fact that these different categories also belong to the category of sensitive documents, the purpose of preventing the leakage of sensitive documents can still be achieved.

6. Conclusion

This paper introduces intelligent classification technology to realize automatic classification and gradation of sensitive data of enterprises. Through the intelligent management of enterprise data, enterprises can quickly grasp the specific quantity and distribution of the information held by the enterprise, greatly reducing the learning cost of the system users and improving the efficiency of work and the accuracy of data classification and gradation. The enterprise data classification and gradation system compensates for the problem of insufficient technical support and system adaptability of the diversified system of sensitive data classification and management through the integration of AI intelligent classification technology. But, when referring to the enterprise document classification accuracy, there is still some misclassification. Inaccurate classification will affect the accuracy of detection of sensitive documents. In the future, the accuracy of document classification should be improved in the classification and gradation system, and the misclassification should be minimized to the greatest extent to provide a more accurate pretechnical means for the enterprise data security protection and data leakage prevention.

Data Availability

The data used to support the findings of this study are included within the article.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported by National Natural Science Foundation (61827811), National Defense Basic Research Program (JCKY2019407C002), the Hebei Provincial Education Departments Support Plan (SLRC2019042), the Hebei Province Funding Project for the Introduction of Overseas Students (C20200364), and China National Petroleum Corporation Information Technology Construction Project (CNPC-IT-2018-N001).

References

[1] J. Campos, P. Sharma, U. G. Gabiria, E. Jantunen, and D. Baglee, "A big data analytical architecture for the asset management," *Procedia CIRP*, vol. 64, pp. 369–374, 2017.

[2] W. Song, Y. Zhang, J. Wang, H. Li, Y. Meng, and R. Cheng, "Research on characteristics and value analysis of power grid data asset," *Procedia Computer Science*, vol. 139, pp. 158–164, 2018.

[3] Z. Liu, B. Hu, B. Huang et al., "Decision optimization of low-carbon dual-channel supply chain of auto parts based on smart city architecture," *Complexity*, vol. 2020, no. 5, 14 pages, Article ID 2145951, 2020.

[4] T.-M. Choi, H. K. Chan, and X. Yue, "Recent development in big data analytics for business operations and risk management," *IEEE Transactions on Cybernetics*, vol. 47, no. 1, pp. 81–92, 2017.

[5] V. V. Glukhov, I. V. Ilin, and A. B. Anisiforov, "Problems of data protection in industrial corporations enterprise architecture," in *Proceedings of the 8th International Conference on Security of Information and Networks*, pp. 34–37, Xi'an, China, 2015.

[6] D. Core, "Applications of text classification to enterprise support documents," *Dissertations*, 2012.

[7] S. Aier, C. Riege, and R. Winter, "Classification of enterprise architecture scenarios-an exploratory analysis," *Enterprise*

Modelling and Information Systems Architectures (EMISAJ), vol. 3, no. 1, pp. 14–23, 2008.

[8] K. Höne and J. H. P. Eloff, "Information security policy - what do international information security standards say?" *Computers & Security*, vol. 21, no. 5, pp. 402–409, 2002.

[9] S. Alneyadi, E. Sithirasenan, and V. Muthukkumarasamy, "A semantics-aware classification approach for data leakage prevention," *Australasian Conference on Information Security & Privacy*, vol. 8544, 2014.

[10] H.-C. Yan, J.-H. Zhou, and C. K. Pang, "Gaussian mixture model using semisupervised learning for probabilistic fault diagnosis under new data categories," *IEEE Transactions on Instrumentation and Measurement*, vol. 66, no. 4, pp. 723–733, 2017.

[11] M. Parasol, "The impact of China's 2016 Cyber Security Law on foreign technology firms, and on China's big data and Smart City dreams," *Computer Law & Security Review*, vol. 34, no. 1, pp. 67–98, 2018.

[12] T. Lewellen, G. J. Silowash, D. L. Costa et al., *Insider Threat Control: Using Plagiarism Detection Algorithms to Prevent Data Exfiltration in Near Real Time*, Carnegie Mellon University, Pittsburgh, PA, USA, 2011.

[13] M. Sogodekar, S. Pandey, I. Tupkari, and A. Manekar, "Big data analytics: hadoop and tools," in *Proceedings of the 2016 IEEE Bombay Section Symposium (IBSS)*, pp. 1–6, IEEE, Mumbai, India, 2016.

[14] H. Lyu, P. Li, R. Yan, H. Qian, and B. Sheng, "High-availability deployment for large enterprises," in *Proceedings of theInternational Conference on Progress in Informatics & Computing (PIC)*, pp. 503–507, IEEE, Shanghai, China, 2016.

[15] S. Daskalaki, I. Kopanas, and N. M. Avouris, "Predictive classification with imbalanced enterprise data," *Recent Advances in Data Mining of Enterprise Data: Algorithms and Applications*, vol. 6, pp. 147–188, 2008.

[16] E. Nwafor, P. Chowdhary, and A. Chandra, "A policy-driven framework for document classification and enterprise security," in *Proceedings of the 2016 Intl IEEE Conferences on Ubiquitous Intelligence & Computing, Advanced and Trusted Computing, Scalable Computing and Communications, Cloud and Big Data Computing, Internet of People, and Smart World Congress*, pp. 949–953, Toulouse, France, 2016.

[17] H. Liang, J. Zou, K. Zuo, and M. J. Khan, "An improved genetic algorithm optimization fuzzy controller applied to the wellhead back pressure control system," *Mechanical Systems and Signal Processing*, vol. 142, Article ID 106708, 2020.

[18] H. Liang, A. Xian, M. Mao, P. Ni, and H. Wu, "A research on remote fracturing monitoring and decision-making method supporting smart city," *Sustainable Cities And Society*, vol. 62, Article ID 102414, 2020.

[19] D. Ben-David, T. Domany, and A. Tarem, "Enterprise data classification using semantic web technologies," in *Proceedings of the Semantic Web-ISWC-International Semantic Web Conference*, 2010.

[20] H. Liang, J. Zou, Z. Li, M. J. Khan, and Y. Lu, "Dynamic evaluation of drilling leakage risk based on fuzzy theory and PSO-SVR algorithm," *Future Generation Computer Systems*, vol. 95, pp. 454–466, 2019.

[21] A. V. Savchenko, *Search Techniques in Intelligent Classification Systems*, Springer International Publishing, Berlin, Germany, 2016.

[22] Z. Huang, X. Xu, J. Ni, H. Zhu, and C. Wang, "Multimodal representation learning for recommendation in Internet of things," *IEEE Internet of Things Journal*, vol. 6, no. 6, pp. 10675–10685, 2019.

[23] A. Abu-Errub, "Arabic text classification algorithm using TFIDF and chi square measurements," *International Journal of Computer Applications*, vol. 93, no. 6, pp. 40–45, 2014.

[24] Y. Zhang, R. Zhu, Z. Chen, J. Gao, and D. Xia, "Evaluating and selecting features via information theoretic lower bounds of feature inner correlations for high-dimensional data," *European Journal of Operational Research*, 2020.

[25] K. A. Keating and S. Cherry, "Use and interpretation of logistic regression in habitat-selection studies," *Journal of Wildlife Management*, vol. 68, no. 4, pp. 774–789, 2004.

[26] S. B. Kim, K. S. Han, H. C. Rim, and S. H. Myaeng, "Some effective techniques for naive Bayes text classification," *IEEE Transactions on Knowledge & Data Engineering*, vol. 18, no. 11, pp. 1457–1466, 2006.

[27] S. Li, S. Ding, and L. Qian, "The decision tree classification and its application research in land cover," *Remote Sensing Technology & Application*, vol. 17, no. 1, 2002.

[28] M. Potamias, F. Bonchi, A. Gionis, and G. Kollios, "K-nearest Neighbors in uncertain graphs," *Proceedings of the VLDB Endowment*, vol. 3, no. 1-2, pp. 997–1008, 2010.

[29] J. Lehečka and J. Švec, "Improving multi-label document classification of czech news articles," *Text, Speech, and Dialogue*, Springer International Publishing, Berlin, Germany, pp. 307–315, 2015.