AN OPTIMIZED METHOD FOR MASSIVE SENSITIVE DATA CLASSIFICATION IN AN INDUSTRY ENVIRONMENT

Qi Zhong, Shichang Gao and Bo Yi

Department of Computer Science and Engineering,
Northeastern University, Shenyang, Liaoning, China

ABSTRACT

In the era of big data, data is endowed with higher potential value. However, new challenges are also brought to data security, especially for the sensitive data in an industrial environment. Nowadays, with the development of industrial internet, enterprises connect each other, under which a slight carelessness may lead to the leakage of sensitive data, which will bring inestimable losses to enterprises. Hence, sensitive data classification is required as a secure way to avoid such situation. This paper presents a sensitive data classification method based on an improved ID3 decision algorithm. Firstly, we introduce the idea of attribute weighting to optimize the basic structure of traditional ID3. Secondly, we use the weighted information gain to select nodes during tree construction, which improves multi-value bias defect compared with the traditional algorithm. Experimental results show that we can achieve branching accuracy up to 97.38%.

KEYWORDS

Sensitive data, Data classification, ID3 decision tree, Industrial environment.

1. INTRODUCTION

Since entering the 21st century, the importance of data has become increasingly apparent and the amount of data has grown exponentially. Such large-scale data promotes the rapid development of the Internet, artificial intelligence and other emerging technologies. With the continuous progress of science and technology, the era of big data has arrived. Although big data brings a lot of convenience, the huge information base also brings some difficulties for accurate screening of information. Therefore, effective classification of big data is an important mean to improve the efficiency of data processing in the network big data platform, as well as an available way to improve users' query experience. Data classification can improve data governance and facilitate data management.

The era of Industry 4.0 is known as "Fourth Industrial Revolution", commonly understood as the transformation from traditional manufacturing to intelligent manufacturing [1]. Industrial Internet is a technology that uses Internet of Things for large-scale industrial manufacturing. The idea behind industrial internet is based on big data, which can process and transmit information more accurately, saving costs for large-scale manufacturing enterprises and achieving optimal output efficiency. Large-scale manufacturing industry chain is a complex network chain consisting of a series of different types of enterprises. Each enterprise, as a data source, continuously generates massive and heterogeneous data at each stage of industrial chain, which is of great value in industrial chain management. There are many enterprises and departments

David C. Wyld et al. (Eds): NLPA, AIS, BDAP, SOENG, IPPR - 2022
pp. 49-62, 2022. CS & IT - CSCP 2022 DOI: 10.5121/csit.2022.121405

involved in industrial chain network, and both cooperative and competitive relations are required among them, which brings challenges to data sharing among enterprises.

On the one hand, all stages of the industrial chain urgently need transparent access and sharing of key information, such as production information, to solve the problem of information asymmetry between upstream and downstream enterprises, improve enterprise cooperation efficiency and reduce costs. On the other hand, there is competition among core enterprises of the industrial chain, such as suppliers at the same level. For the protection of their own privacy, enterprises do not want private data such as trade secrets and financial information to be obtained by others. How to ensure transparent access and sharing of necessary data among enterprises and realize protection of sensitive data is an urgent problem to be solved.

Generally, the data classification needs labels and categories for a given data type. These types will be used to set sensitivity and confidentiality levels. Sensitive data refers to data that is unknown to public, has actual or potential use value, and will cause harm to society, enterprises or individuals if lost, improperly used or unauthorized access [2]. With the development of big data, Data security faces new challenges. After years of infomationization construction, large amounts of enterprise-level data has been accumulated by large-scale manufacturing industry chain, including customer data, marketing data, production data and other important sensitive data.

However, with the large-scale use of network data, especially in the application of big data, there are loopholes in the big data platform itself, and a slight carelessness may lead to data leakage. Once these sensitive data are leaked, it will inevitably bring incalculable losses to enterprises. If all data are set to high sensitivity level instead of data classification, it will lead to high cost, high operational complexity and high cost in data security protection, and also greatly reduce enterprise cooperation efficiency. However, if all data are set to low sensitive level, due to leakage of important private data, these data will be obtained by attackers or used by malicious people to engage in business transactions or steal personal privacy, thus causing unpredictable losses to interests of enterprise and employees. Therefore, sensitive data should be properly graded and classified as needed.

Existing sensitive data classification methods focus on classification of imbalanced data (e.g. [3]-[5]) and feature selection (e.g. [6]-[8]). At present, existing imbalanced data classification technology can be roughly divided into data-level method and algorithm-level method [3]. Datalevel strategy is more general because it does not rely on classifier model; However, when creating samples for minority classes [4] are added, valuable information may be eliminated due to the loss of majority class samples, leading to over fitting. Algorithm-level strategy is not widely used because it is limited by specific classifiers or datasets [3]. The existing imbalanced data classification models mainly focus on the classification performance of majority class samples, which does not correspond to the fact that the sensitivity of sensitive data is inversely proportional to the scale of data. Techniques such as [5] balance different classes by enlarging samples with minority classes (oversampling) or discarding samples from majority classes (under sampling), achieving improved data-level sampling. Although undersampling technique equalizes different categories of samples and reduces time cost, the loss of important information in sensitive data with imbalanced scale will appear.

Feature selection methods ([6]-[8]) can effectively project high-dimensional data into relatively low-dimensional space. Feature selection is mostly based on filtering and wrappering approaches that suffer from poor classification accuracy, high computational cost, irrelevant selection and redundant features [6]. This is due to the limitations of adopted objective functions leading to overestimation of feature significance. On the contrary, hybrid feature selection methods based on information theory and nature-inspired meta-heuristic algorithms have advantages of high computational efficiency, good scalability, avoidance of redundancy and less informative features, and independence from classifier. However, these methods have three common shortcomings on sensitive data classification: *(1) poor trade-off between exploration and exploitation phase; (2) trapping in optimal local solution; (3) avoiding irrelevance and redundancy of selected features.*

To address the above challenge, we propose an improved ID3 decision tree based efficient sensitive data classification method. Specifically, we divide sensitive data into four sensitive levels according to actual situation, use data subset with sensitivity label to train the model, and then classify sensitive level of data by integrating characteristics with this model. Based on traditional ID3 algorithm, we introduces the idea of attribute weighting and uses the weighted information gain to select nodes during tree construction, which improves multi-value bias defect in traditional algorithm. We use different labels and training set size, and add error sensitive data of manual intervention for training to prevent the occurrence of over-fitting phenomenon. Our research makes full use of the information in sensitive data and minimizes the loss of information while preventing irretrievable loss caused by the leakage of important privacy data, which has great practical significance. Using ID3 decision tree can achieve our goal well. The model can not only keep concise structure, but also synthesize various data characteristics to decide the final classification of data, whose accuracy meets industrial requirements. The experimental results indicate that the proposed method can ensure maximum accuracy while avoiding the problem of low model efficiency caused by multi-value bias phenomenon compared with existing methods.

The rest of this work are as follows: Section 2 summarizes the related work about the sensitive data classification. Section 3 introduces the proposed improved ID3 decision tree algorithm for sensitive data classification. Section 4 evaluates the performance of proposed decision tree model. And section 5 concludes this work.

2. RELATED WORK

There are three main kinds of technologies used for sensitive data classification, which are differential privacy protection, quantum machine learning and neural network. We present them as follows.

2.1. Differential Privacy Protection based Sensitive Classification

Dwork et al. [9] put forward "differential privacy" protection framework in 2017, injecting an appropriate amount of random noise into original data sets, processed and deformed data sets, query results and query functions of statistical functions to ensure that whether the target data is stored in database does not affect search results of functions. At the expense of certain prediction accuracy, data privacy is not compromised.

Private Aggregation of Teacher Ensembles (PATE) strategy proposed by Nicolas Papernot et al. [10], combined with differential privacy protection framework, proposes a solution to the problem of data privacy leakage in classification task of image data sets with few categories, and protects privacy security of sensitive attribute values of specific data in classification task. However, this method is not satisfactory for classification task of image data sets with a large number of categories, and can not achieve the function of balancing model accuracy and data privacy security.

Uri Stemmer [11] proposes to apply the method of local differential privacy in k-means clustering algorithm to adapt to privacy needs of different users, reduce concerns about exposure of specific sensitive attribute values. Although local privacy budget is considered in this method, the overall differential privacy budget of the algorithm is not taken into account, which will lead to lower accuracy of new data partition results of the trained model. Literature [12] proposes a clustering scheme LDPK-modes (Local Differential Privacy K-modes) that can support localized differential privacy technology. Compared with traditional privacy protection algorithm based on centralized differential privacy clustering, it solves the problem that it is difficult to find a trusted third party to collect and process sensitive data in practical application scenarios, reduces the possibility of the third party stealing user sensitive data, and considers the possibility of user sensitive data leakage from another perspective. However, this method does not guarantee the accuracy of exploring laws and properties of data because clustering is sensitive to initial center point.

2.2. Quantum Machine Learning based Sensitive Classification

Existing quantum machine learning privacy protection mainly focuses on the following three types of privacy protection technologies: privacy protection technology based on differential privacy, privacy protection technology based on homomorphic encryption and privacy protection technology based on secure multi-party computing.

Quantum differential privacy is a quantum scheme of differential privacy, which has been used to protect the privacy of quantum machine learning models such as classification, regression and neural network. Du et al. [13] proposes a differential privacy protection scheme for quantum classifiers by using depolarization noise of quantum channels, and gives robustness bounds of anti-disturbance. Watkins et al. [14] proposes a quantum machine learning algorithm based on differential privacy framework to realize privacy protection through differential privacy optimization algorithm based on Gaussian noise.

Homomorphic encryption is widely used in quantum computing to realize privacy-protecting quantum computing. Gong et al. [15] proposes a quantum homomorphic encryption method based on improved T-gate update, and realized a quantum K-means algorithm based on cloud server for privacy protection. The increase in the number of H-gate and T-gate in algorithm leads to an increase in algorithm complexity and a decrease in accuracy.

Quantum secure multi-party computing is a quantum solution for secure multi-party computing, which has been used in privacy protection research of quantum machine learning. Li et al. [16] proposes a private single-party delegation training protocol for variable component sub classifiers based on blind quantum computing, and extends this protocol to multi-party private distributed learning with differential privacy. Xia et al. [17] proposes QuantumFed, a quantum federated learning framework, which enables multiple quantum nodes with local quantum data to train model together. Chehimi et al. [18] proposes a complete quantum federated learning framework for clients with quantum computing capabilities.

2.3. Neural Network based Sensitive Classification

Deep learning prediction model based on privacy protection was first proposed by Dowlin et al. [19]. This model implements CryptoNets, a protocol for privacy protection of convolutional neural network. Mohassel et al. [20] adopts a hybrid protocol framework to realize training of neural networks and evaluate prediction classification of both parties. Liu et al. [21] proposes MiniONN, which uses single-instruction multi-data technology (SIMD) to transform existing neural network into observable neural network.

Some studies have also focused on privacy of data in training stage of neural networks, mainly aiming at back propagation algorithms. Bu et al. [22] proposes a back propagation algorithm for privacy protection based on BGV encryption mechanism in the cloud, which uploads all computing load to the cloud to reduce computing overhead, and also uses BGV [23] to protect data privacy during processing. Zhang et al. [24] proposes to use BGV encryption scheme to effectively support secure computing of high-order back propagation algorithm, so as to conduct model training in the cloud. In order to reduce the depth of multiplication, after each iteration, the updated weights are sent back to all parties for decryption and re-encryption, so the communication cost of this scheme is very high.

Most of the existing work above focus on privacy protection of image data, classifier data or sensitive data classification in training process, while not suitable for enterprise-level sensitive data generated in practical large-scale manufacturing industry chain. In addition, compared with above methods, proposed sensitive data classification method maintains a very concise algorithm structure while ensuring accuracy in line with industry requirements, and is still equipped with the ability to classify data by integrating various features, which has great practical significance.

3. Improved ID3 Decision Tree for Sensitive Data Classification

Decision tree algorithm is one of the most basic and effective algorithms in artificial intelligence and machine learning. Decision tree algorithm constructs mathematical analysis model and obtains basic classification rules through training and mining rule of irregular and disordered sample data, and then makes prediction classification of target data set. On the basis of the preprocessing sensitive data, the ID3 decision tree is constructed and optimized in this paper to classify the sensitive industrial data. Due to the simple ideas and strong learning ability of ID3 for data sample features, we first adopt it as the basis for data classification, which is then improved by using the attribute weighting. After that, the improved ID3 algorithm is used to achieve more accurate sensitive data classification.

3.1. Basic ID3 Description

The basis of ID3 algorithm looks for the attribute with the largest amount of information in input data set as a node of decision tree, and then classifies sample data according to different attribute values under this attribute. The core of the algorithm is information entropy, which is defined as follows:

$$Ent(D) = -\sum_{i=1}^{m} p_i \log_2(p_i) \tag{1}$$

where p_i is the probability of the target attribute appearing in sample data set *D*.

Information entropy is used to calculate the purity of a sample set, which represents the number of sample types contained in this node. It can be seen from Formula (1) that higher information entropy represents lower sample purity and more sample categories of this node. On the contrary, if information entropy is lower, the purity of this node sample will be higher, and sample data of this node will have fewer categories.

Information gain is the change value of information entropy of sample data before and after attribute division, which can be used to measure difference of information entropy in probability distribution. The information gain divided according to attribute *A* is defined as:

$$Gain(A) = \mathrm{Ent}(D) - Ent_A(D) \tag{2}$$

ID3 algorithm calculates information gain of each attribute through information entropy and takes it as the dividing criterion to compare which attribute is selected for node splitting. The attribute with the highest information gain value is selected as the standard for each partition, and the process is repeated until information entropy of the sample set is zero. ID3 algorithm is simple and feasible with intuitive and clear process. It can train classification model well for most discrete data.

3.2. Improved ID3 Algorithm based on Attribute Weighting

In ID3 algorithm, each non-leaf node selects the attribute with the maximum information gain value as split attribute, which often ignores the attribute that is very important for decision making but has few values, namely the defect of multi-value bias in ID3 decision tree. Aiming at this shortcoming, this paper proposes an improved ID3 algorithm based on attribute weighting. In this algorithm, attribute weights are introduced in the calculation of information gain, and correlation coefficients are used as attribute weight coefficients to ensure rationality of weighted coefficients.

Correlation coefficient reflects closeness of the relationship between random variables, and its value is between -1 and 1. In the same sample space, the closer the absolute value of correlation coefficient between attributes is to 1, the higher the correlation degree between attributes is, so the correlation weight between attributes can be calculated by using the meaning of correlation coefficient. Set A_c as a conditional attribute and A_d as a decision attribute, and the correlation coefficient between attribute A_c and attribute A_d can be calculated according to Formula (3).

$$\rho(A_c, A_d) = \frac{\operatorname{cov}(A_c, A_d)}{\sqrt{D(A_c)D(A_d)}} \tag{3}$$

where $cov(A_c, A_d)$ is the covariance of A_c and A_d, $D[A_c]$ is the variance of A_c, and $D[A_d]$ is the variance of A_d.

Thus, the correlation weight of attribute A_c to decision attribute A_d can be obtained:

$$\lambda = |\rho(A_c, A_d)| = \left| \frac{\operatorname{cov}(A_c, A_d)}{\sqrt{D(A_c)D(A_d)}} \right| \tag{4}$$

Therefore, according to the characteristics of ID3 algorithm that selects the maximum information gain as splitting basis and the significance of correlation coefficient, Formula (5) is obtained after the improvement of attribute weighting from Formula (2) :

$$Gain(A) = \mathrm{Ent}(D) - (1-\lambda)Ent_A(D) \tag{5}$$

3.3. Sensitive Data Classification Based on Improved ID3 Algorithm

The sensitive data classification based on improved ID3 algorithm proposed in this paper is mainly composed of three parts: data preprocessing, feature selection of decision tree nodes and decision tree construction.

3.3.1. Data Preprocessing

There are many non-numeric attribute values in sensitive data. In order to make data more suitable for the model to be established, and improve computing efficiency of ID3 decision tree, the specific operation of data conversion in this paper is to transform non-numerical characteristic parameters in original sensitive data into numerical characteristic parameters. For example, in original sensitive data sample set, the characteristic value of *safety* feature in the data is text type, with four levels of "*Low*", "*Med*", "*High*" and "*Vhigh*". Convert them correspondingly to numeric types, that is, "*Low*"—1, "*Med*"—2, "*High*"—3, and "*Vhigh*"—4 respectively, which then achieves the operation of data type conversion.

3.3.2. Feature Selection

The feature selection criterion of each decision node in ID3 decision tree algorithm is based on information gain theory. According to the improved ID3 algorithm proposed in this paper, feature with the maximum information gain after improvement is selected as the feature of current node by calculating information gain of each feature based on attribute weighting.

According to Formula (1), (3) and (5), taking unsegmented training sample set of sensitive data as an example, information entropy, weighting coefficient and improved information gain of six conditional features of sensitive data are calculated as shown in Table 1.

Table 1. Key calculation results of unsegmented training sample set.

Feature	Information Entropy	Weighting Coefficient	Information Gain
buying	1.10	0.28	0.40
maint	1.13	0.24	0.35
doors	1.20	0.05	0.06
persons	0.99	0.34	0.55
lug-boot	1.17	0.04	0.07
safety	0.94	0.41	0.64

According to data in Table 1, safety feature has the maximum information gain after attribute weighting, so safety feature is selected as the root node feature. The selection of other node features is similar to the above process.

3.3.3. Decision Tree Construction

The core stage of decision tree algorithm is tree construction process. Suppose training set D have M attributes A_i, where i = 1, 2, ..., M. For the improved ID3 algorithm proposed in this paper, its tree-construction process is as follows:

Step 1. Construct a node. If all samples are on this node, stop the algorithm, change the node into a leaf node, and mark it with this class.

Step 2. If all samples are not on the same node, work out information entropy of each attribute, figure out improved information gain of each attribute according to Formula (5), and then select the attribute with maximum information gain as classification basis. That is, *best_attribute* = A_j where A_j satisfies Formula (6):

$$Gain(A_j) = \max_{1 \le i \le M}(\mathrm{Ent}(D) - (1-\lambda)Ent_{A_i}(D)) \tag{6}$$

Step 3. Assume that the attribute value of A_j has V different discrete values, create a branch for each value, and divide sample set D into V subsets $\{D_1, D_2, ..., D_v\}$ according to attribute A_j.
Step 4. Repeat ***Step 1-Step 3*** until all attributes are used up.

In the process of tree construction, improved ID3 algorithm excludes the special case of ***Step 1*** and judges all attribute values in attribute set. If there are multiple values for multiple attributes in the attribute set, the optimal attribute is selected, and sub-nodes are divided according to different values of sample data on the optimal attribute. After dividing data of the current node, the features in data set that have been divided in the previous step are removed to generate data subset, and a new round of division is continued. Until data set cannot be divided, the decision tree stops growing, and finally a complete classification decision tree is returned.

Table 2. Pseudo code of improved ID3 decision tree construction.

```
Algorithm Improved ID3 Decision Tree Construction

Input Training Data Set; Attribute List
Initialize a node $N$
if samples are on this node
  return $N$ as a leaf node and mark $N$ with this class else
    for $i \in \{1,2,...,M\}$ do
        figure out information gain with attribute weight of each
        attribute $A_j \in \{A_1,A_2,...,A_{M-i}\}$ in Attribute List select
        best_attribute with maximum information gain remove
        best_attribute from Attribute List mark $N$ as
        best_attribute
       for attribute value $a_k \in \{a_1,...,a_v\}$ in best_attribute do
          create a branch and divide the samples. end
    for
    end for end if
Output A Decision Tree
```

3.3.4. Sensitive Data Classification Based On ID3 Improved Algorithm

The improved ID3 algorithm proposed in this paper introduces attribute weights in calculation of information gain, considering the correlation between attributes, which not only makes up for multi-value bias defect of ID3 algorithm to a certain extent, but also applies to data analysis in more cases. The specific process of sensitive data classification based on the improved ID3 algorithm proposed in this paper is described as follows:

Step 1. Divide sensitive data set into training set and test set in a ratio of 7:3.

Step 2. Calculate the information entropy and attribute weight of each condition attribute of sensitive data in training set, and introduce weights into the calculation of information gain according to Formula (7) :

$$Gain(A_i) = \mathrm{Ent}(D) - (1 - \left| \frac{\mathrm{cov}(A_i, A_d)}{\sqrt{D(A_i)D(A_d)}} \right|) Ent_{A_i}(D) \tag{7}$$

where i = 1, 2, ... , M -1.

Step 3. Select the attribute with the maximum information gain after improvement as the split attribute to construct a decision tree.

Step 4. Test the constructed decision tree with test set.

The specific flow chart of sensitive data classification based on improved ID3 algorithm is shown in Figure 1.

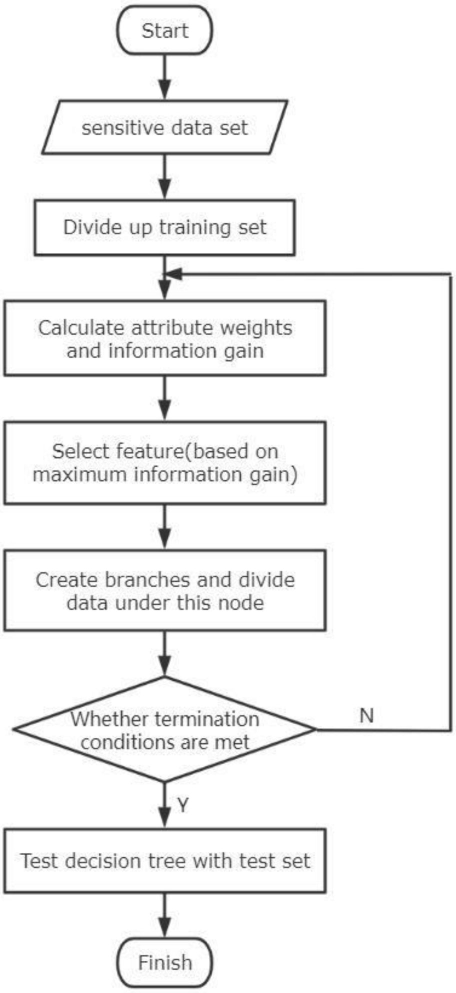
Start
sensitive data set
Divide up training set
Calculate attribute weights and information gain
Select feature(based on maximum information gain)
Create branches and divide data under this node
Whether termination conditions are met
N
Y
Test decision tree with test set
Finish

Figure 1. Algorithm flow chart

4. Performance Evaluation

4.1. Setup

A total of 3387 sensitive data in data set are used in the experiment, and each piece of data has 7 features, including 6 conditional features and 1 decision feature (sensitivity level). Sensitive data are divided into four levels: public data, external sensitive data, internal inter-departmental sensitive data and internal intra-departmental sensitive data. Each level has 2353,680,156 and 198 pieces of data respectively.

The sensitive data set consists of training data set and test data set. Training sample set constructs ID3 decision tree classification model while test sample set evaluates the performance of ID3 decision tree model. Sample selection should be representative, typical and complete. In order to improve classification accuracy, the data set is divided into training set and test set at the ratio of 9:1, 8:2, 7:3 and 6:4, respectively. The experimental results are shown in Figure 1. After experiments on training sets of different scales, it is finally found that the optimal classification effect of decision tree model can be achieved by dividing training data set according to the ratio of 7:3. After dividing by 7:3, there are 2370 sensitive data pieces in training set and 1017 sensitive data pieces in test set.

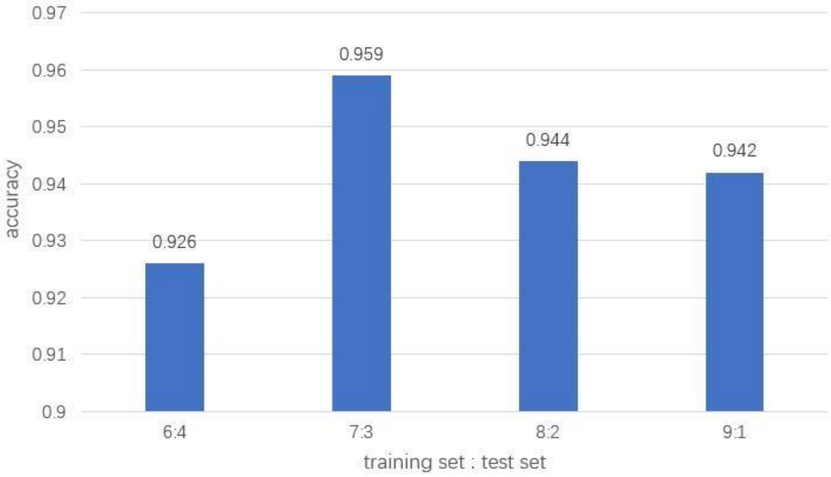

Figure 2. Branching accuracy under different training set scale

4.2. Metric

Generally, the evaluation of decision tree classification model is mainly carried out from the following two aspects:

(1) Accuracy: Accuracy refers to the ability of decision tree classification model to accurately predict category of new or unknown target data. The accuracy of prediction is related to the number of samples in training set, the number of decision attributes, the number of attribute values and the distribution of samples to be predicted.

(2) Robustness: Robustness refers to the ability of the classification model to accurately classify data sets in the case of incorrect data and incomplete data. Data collected in real life will inevitably have incomplete data, data errors, data redundancy and other phenomena, so the constructed decision tree classification model should have the ability to

deal with these data well, so as to avoid the existence of such data affecting performance of classifier.

4.3. Result Analysis

4.3.1. Accuracy Analysis

Table 3 shows the accuracy of data classification of each sensitivity level. As can be seen from Table 3, the classification results of data of all sensitivity levels are highly accurate, which can meet the needs of industry.

Table 3. Accuracy of different sensitivity level data.

Sensitive level	public	external	Inter-departmental	Intra-departmental
Sample number	2353	680	156	198
Error number	47	40	9	14
Accuracy	98.0%	94.1%	94.2%	92.9%

Figure 3 shows the classification accuracy of traditional ID3 decision tree algorithm and improved ID3 decision tree algorithm in this paper under different training sample set sizes. As can be seen from Figure 3, the accuracy of ID3 decision tree algorithm after attribute weighting is significantly improved. And the smaller the training set size, the better the performance of improved ID3 algorithm compared with traditional one.

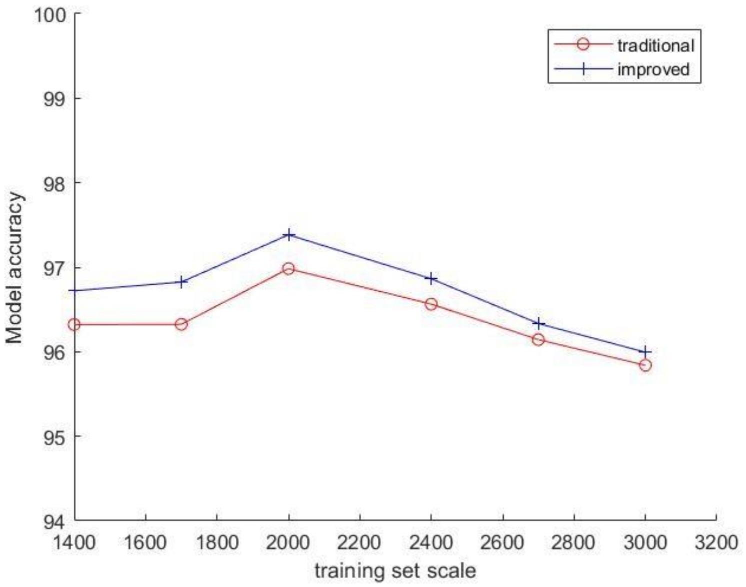

Figure 3. Accuracy in different training set scale

4.3.2. Robustness Analysis

In order to verify the robustness of the improved ID3 decision tree algorithm in this paper, noise data of manual intervention is added for training under the same data set size. The experimental results are shown in Figure 4. As can be seen from Figure 4, when the proportion of noise data is less than 8%, the robustness of improved ID3 decision tree algorithm is significantly higher than that of traditional ID3 decision tree algorithm. However, when the proportion of error sensitive data is 9% and 10%, the robustness of traditional ID3 algorithm is better than that of improved ID3 algorithm. The reason for this phenomenon may be that the calculation of attribute weights is deviated due to excessively high noise data, which affects the selection of feature in each node and worsens classification results.

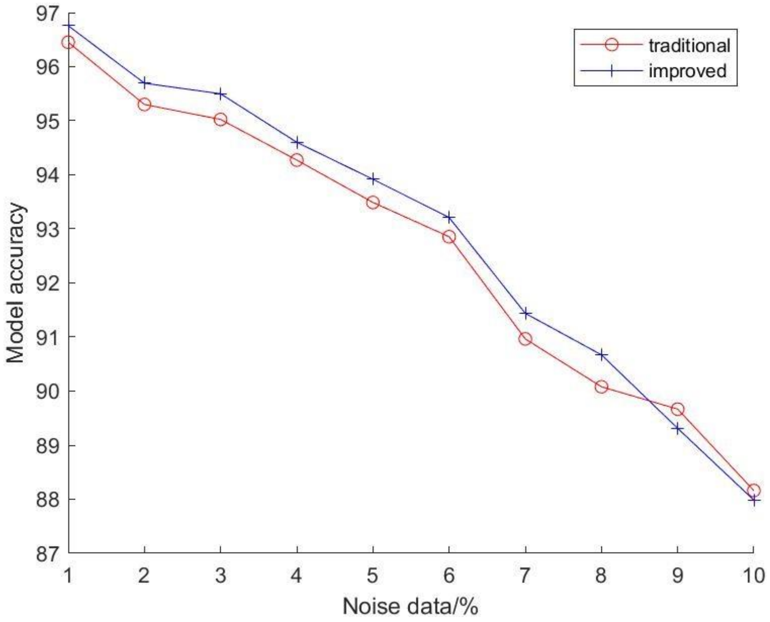

Figure 4. Accuracy under different noise data ratio

In addition, we also verify robustness of the algorithm from the perspective of data sources. The Vote, DNA-data, Weather and Soybean data sets are selected for training. The experimental results are shown in Figure 5. As can be seen from Figure 5, under four different data sets, the improved ID3 algorithm proposed in this paper has stable classification performance, that is, good robustness. Compared with the traditional method, the improved algorithm also has higher accuracy, which further verifies the performance optimization effect of this algorithm.

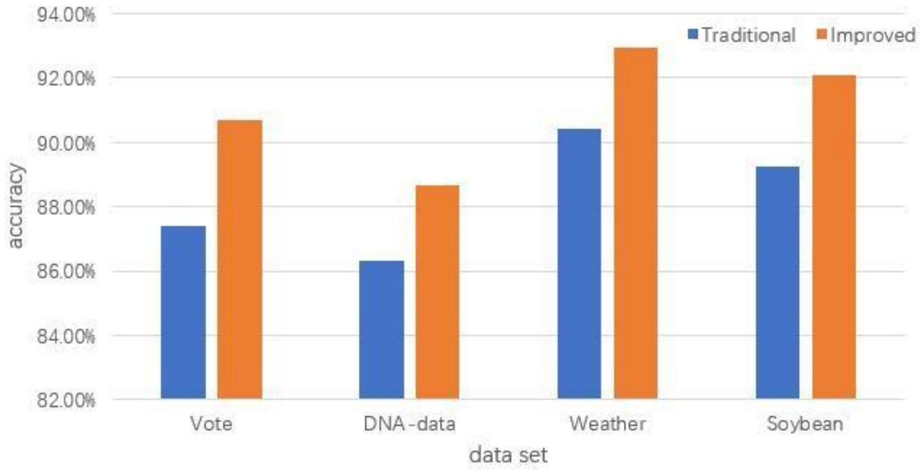

Figure 5. Accuracy under different data set

5. CONCLUSION

The sensitive data classification based on improved ID3 decision tree algorithm by this invention realizes the maximum utilization of sensitive data generated by large-scale manufacturing industry chain in big data environment. While making full use of information in sensitive data, it prevents irreparable loss caused by the leakage of important sensitive data, which is of great practical significance. According to experimental results, the branch accuracy of improved ID3 decision tree algorithm for sensitive data can reach 98%, which is significantly higher compared with traditional ID3 decision tree. In addition, it can still meet industrial needs with good test results even in the case of artificial error data added.

ACKNOWLEDGEMENTS

This work is supported by the National Key R&D Program of China under Grant No. 2021YFB3300900.

REFERENCES

[1] Miguel Oliveira & Daniel Afonso, (2019) “Industry Focused in Data Collection How Industry 4.0 is Handled by Big Data[C]”, *Proceedings of 2019 International Conference on Data Science and Information Technology*（*DSIT 2019*）, DOI: 10.26914/c.cnkihy.2019.092721.

[2] Ji-sung Park, Gun-woo Kim & Dong-ho Lee, (2020) “Sensitive Data Identification in Structured Data through GenNER Model based on Text Generation and NER[C]”, *Proceedings of 2020 2nd International Conference on Computing, Networks and Internet of Things*（*CNIOT 2020*）, DOI: 10.26914/c.cnkihy.2020.033027.

[3] Xia Shuyin & et al, (2021) “Granular Ball Sampling for Noisy Label Classification or Imbalanced Classification[J]”, *IEEE transactions on neural networks and learning systems*, pp112

[4] Sebastián Maldonado & Julio López, (2018) “Dealing with high-dimensional class-imbalanced datasets: Embedded feature selection for SVM classification[J]”, *Applied Soft Computing*, 67: pp94-105.

[5] Funa Zhou & et al, (2020) “Deep learning fault diagnosis method based on global optimization GAN for unbalanced data[J]”, *Knowledge-Based Systems*, 187(C): pp104837-104837.

[6] Andrea Bommert & et al, (2020) “Benchmark for filter methods for feature selection in highdimensional classification data[J]”, *Computational Statistics and Data Analysis*, 2020, 143(C): pp106839-106839.

[7] Li Gui & et al, (2020) “DLEA: A dynamic learning evolution algorithm for many-objective optimization[J]”, *Information Sciences*, 574: pp567-589.

[8] Sadeghian Zohre, Akbari Ebrahim & Nematzadeh Hossein, (2021) “A hybrid feature selection method based on information theory and binary butterfly optimization algorithm[J]”, *Engineering Applications of Artificial Intelligence*, p97

[9] Cynthia Dwork & et al, (2017) “Calibrating Noise to Sensitivity in Private Data Analysis[J]”, *Journal of Privacy and Confidentiality*, 7(3): pp17-51.

[10] Nicolas Papernot, Shuang Song, Ilya Mironov & et al, (2018) “Scalable private learning with PATE[C]”, *The proceedings of the 6th International Conference on Learning Representations*, Vancouver, Canada: ICLR: pp1-15.

[11] Uri Stemmer, (2019) “Locally Private k-Means Clustering[J]”, *CoRR*, abs/1907.02513.

[12] Peng Chunchun, Chen Yanli & Xun Yanmei, (2021) “k-modes clustering guaranteeing local differential privacy[J]”, *Computer Science*, 48(2): pp105-113.

[13] Du Yuxuan & et al, (2021) “Quantum noise protects quantum classifiers against adversaries[J]”, *PHYSICAL REVIEW RESEARCH*, 3(2).

[14] Watkins WM, Chen SYC & Yoo S, (2021) “Quantum machine learning with differential privacy [J]”, *arXiv preprint*, arXiv:2103.06232.

[15] Gong C, Dong Z, Gani A & et al, (2021) “Quantum k-means algorithm based on trusted server in quantum cloud computing”, *Quantum Inf Process* 20, 130.

[16] Li Weikang, Lu Sirui & Deng Dongling, (2021) "Quantum private distributed learning through blind quantum computing [J]", *arXiv preprint,* arXiv: 2103.08403.

[17] Xia Qun, Li Qun & QuantumFed, (2021) "A federated learning framework for collaborative quantum training [J]", *arXiv preprint,* arXiv: 2106.09109.

[18] Chehimi M & Saad W, (2021) "Quantum federated learning with quantum data [J]", *arXiv preprint,* arXiv, 2106.00005.

[19] Dowlin N, Gilad-Bachrach R, Laine K & et al, (2016) "Cryptonets: Applying neural networks to encrypted data with high throughput and accuracy [C]", *Proceedings of the 33nd In-ternational Conference on Machine Learning,* JMLR. org, pp201-210.

[20] Mohassel P & Zhang YP, (2017) "SecureML: A system for scalable privacy-preserving machine learning [C]", *IEEE Sym-posium on Security and Privacy (SP),* IEEE, pp19-38.

[21] Liu J, Juuti M, Lu Y & et al, (2017) "Oblivious neural network predictions via MiniONN transformations [C]", *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Com-munications Security,* ACM, pp619-631.

[22] Bu FY, Ma Y, Chen Z K & et al, (2015) "Privacy preserving back-propagation based on BCV on cloud[C]", I*EEE 17th International Conference on High Performance Computing and Communications, 2015 IEEE 7th International Symposium on Cyberspace Safety and Security, and 2015 IEEE 12thInternational Conference on Embedded Software and Systems, IEEE,* pp1791 - 1795.

[23] Gentry C, Halevi S, Peikert C & et al, (2013) "Ring switching in BGV-style homomorphic encryption [J]", *Journal of ComputerSecurity*, 21(5): pp663 -684.

[24] Zhang QC, Yang LT & Chen ZK, (2016) "Privacy preserving deep computation model on cloud for big data feature learning [J]", *IEEE Transactions on Computers*, 65(5): pp1351 -1362.

AUTHORS

Qi Zhong is studying for a bachelor's degree in college of Computer Science and Technology at North eastern University in Shenyang, China. Her research interests include neural network, AI-enable network and cognitive radio network.

Gaoshichang, male, 21 years old, is studying at North eastern University for a Bachelor of engineering degree in communication engineering. During the undergraduate period, he devoted himself to scientific research competitions and won many awards. During his undergraduate years, his main research direction was big data classification.

Bo Yi received the BS and MS degrees in computer science from the South-Central University for Nationalities, Wuhan, China, in 2012 and 2015, respectively, and the PhD degree in computer science from North eastern University, Shenyang, China, in 2019. He is currently a lecturer with the College of Computer Science and Engineering, North eastern University, Shenyang, China. His research interests include routing and service function chain in SDN, NFV, deterministic networking, and cloud computing.