Research on Encrypted Text Classification Based on Natural Language Processing

Qiuyi Ren
College of computer and information science, Southwest University, Chongqing, 400715, China

Abstract. In reality, data encryption technology is mostly used to protect the security of text data in the network, but when we need to obtain these data, this layer of encryption becomes an obstruction to obtaining data. The general method uses data mining and data decryption to extract effective information. The experimental data in this article selected 20 categories of text information, and obtained a data set with a difficulty of 1 to classify the encrypted text information. In order to classify encrypted text more effectively, this paper studies the method of using the logistic regression model and the LightGBM model algorithm to directly process encrypted text, which can directly extract and classify the text in the encrypted state. Model evaluation results show that LightGBM is more effective. In addition, this article provides a basic framework for the classification of encrypted text based on natural language processing.

1. Introduction

Big data is a hot topic in current world research, which affects people's life. All kinds of research should not only solve the problem of effective information mining, but also assure the security of information [1]. In accordance with information security, various types of data encryption technology will be used to assure the security of information dissemination in the network. In people's opinion, classification is an inductive form of information, and relevant information is grouped into the same category. The current classification research has a relatively high level of practicality, but the times are changing, the faster the information processing speed is, the better, the complexity of text information is also enhanced, especially the proliferation of some popular information, especially online information, so there is a lot of dynamic information. Although automatic text classification can effectively classify new information, it is suitable for today Processing and organizing this large amount of online text information. However, people's demand on the accuracy of search is higher and higher, which urges us to strengthen the effective text classification in the classification technology, so as to provide a better text classification.

Data encryption technology is a kind of special information transformation technology for both sides of communication according to the agreed rules. It is generally composed of plaintext, ciphertext, algorithm and key. Through the specific information technology to process the data, improve the understanding difficulty, ensure that the information can only be obtained by the users authorized with special processing rules, which greatly ensures the security of information. Encryption technology can be simply divided into two common categories, namely symmetric encryption and asymmetric key encryption. Symmetric encryption is a relatively traditional and simple encryption type, which uses the same key when encrypting and decrypting data. Using symmetric encryption technology, the security of data depends on the key, but the security of key management is not guaranteed, which makes this encryption method difficult to implement in application. The obvious difference between asymmetric key encryption and symmetric encryption is that a set of public key and private key systems are used.
The public key is used to encrypt data, and only the corresponding private key can be used to decrypt it. This method is more secure than asymmetric key encryption, but it also has the problem of slow speed [2][3]. When it is necessary to extract or classify encrypted information, data research becomes more and more complex in the face of the rate of change of updated data and the new types of data, which can extract and optimize the effective information hidden in the massive data. Classification becomes crucial. As the most commonly used data analysis method, data mining is applied in many fields through techniques such as classification, optimization, identification, and prediction. Use data mining technology to process and analyze encrypted data, establish an appropriate system, continuously optimize, improve the accuracy of decision-making, and more conducive to the grasp of effective information. The structure of most of the text information data available on the network is unstructured. For this part of the text, a special processing method is needed to transform it, that is, text mining. Text mining refers to the process of extracting unknown parts from a large amount of text data and better organizing this part of information. The difficulty lies in the need to deal with vague and unstructured text data, which covers knowledge and technologies of multiple disciplines.

Machine learning is widely used in encrypted text classification. The classifier with two-level classification method is evaluated by experiments of standard Reuters-21578, cade12 and 20 newsgroups, which shows that the accuracy of classification is improved, the classification effect is improved, and the classification accuracy of machine learning classifier is effectively improved [4]. Data processing tools using Deep Trust Networks (DBNs) are used to distinguish data categories. The experimental results on MNIST (image dataset) and ISOLET (speech dataset) show that the data features are successfully extracted and classified [5]. Using SVM (Support Vector Machine) for training can better extract valuable information from the website, measure the performance of the system through calculation accuracy and recall value, that is, can represent the effect of the system, and finally calculate the single-value measurement value. Select the short messages in the active network news information, after manual classification, use the words in the short messages as features to generate feature vectors, and train them with SVM machine learning technology. The results show that the system provides a higher performance [6]. Vinayakumar et al. proposed long short-term memory network to study the encrypted text by using embedding techniques. The results showed that accuracy of the word level was 0.94 in the dataset of 5-fold cross validation but the accuracy was 0.42 in real-world test data [7-8].

This article mainly studies a method that can directly extract and classify encrypted text. When the information is encrypted, keywords are extracted and classified directly on the encrypted information. The algorithm models used in this article include the logistic regression model and the LightGBM model, using 20 categories of text information for classification testing.

2. Data Introduction

The text information used in this article comes from the kaggle website. The sample is 9589, including the encrypted information and the result of classification. The number of categories to be classified is 20, and the data set with difficulty of 1 is obtained. The encrypted information of a certain news is as follows:

"O8v^10O#to1#/^tv1^s111t01Otaq>ata_1O_tv1a18Odc1Wat0t01-cvvaacvtv1e#1/O#aOPs111:1A^;t1Ao^c#1aOv611s1s111>ada1E8a0e1J1zaT1O_Oza888T1208a]"aOs1s1ULi1}O-v1"1leOA08aOT1;xli/^T1x#0vt0-Ot1wOOoOT;1a1t01;/#aOs11111-zaWT;1O1VcO#H#801o1t01gOa#af1/"v^""

It is mainly included in text processing. When the text needs to be turned into a vector to form a matrix for input into the model for processing, TF-IDF can be used for processing. TF-IDF, namely word frequency-inverse text frequency, is composed of TF and IDF, and is used for text processing to convert text into a vector that can be processed by the model. The importance of a word is proportional to its frequency in the text, which is recorded as TF; it is inversely proportional to its frequency in the corpus, which is recorded as IDF. The calculation method is:

\[ TF-IDF = TF \times IDF \]
Where TF represents word frequency, the TF of the word \( w \) is calculated by dividing the number of times \( w \) appears in the document by the total number of words in the document. IDF represents the frequency of reverse documents. Although some words may be used very frequently in the text, it is not important, that is, the amount of information is small, words like ‘is’, ‘are’, ‘and’, ‘of’, ‘that’ and so on. Take advantage of the fact that these words are also appear frequently in the corpus to reduce their weight. The IDF of the word \( W \) is calculated by dividing the total number of documents in the corpus by the number of documents with the word \( W \) and calculating the logarithm.

![Figure 1. Proportion of each category in the sample](image)

3. Model

3.1 Logistic regression

Logistic regression is a generalized linear regression analysis model. It is actually a classification method. It is often used in machine learning methods to solve binary classification problems. This model assumes that the data obeys the Bernoulli distribution. By maximizing the likelihood function, gradient descent is used to solve the parameters, so as to achieve the purpose of binary classification of the data [8-12]. Its representation form is (1). When the probability of the input \( X \) belonging to the first category is recorded as, generally when the probability is greater than 0.5, the output result is judged as 1, otherwise it is 0. Based on the existing data, a regression formula is established for the classification boundary to classify and predict the possibility of things happening. The logistic regression model is simple to implement, very efficient, and does not require too much calculation.

The calculation only has relation to the number of features during classification, which is convenient for use in big data scenarios. And the output is the probability score of each sample, which can be easily classified. At the same time, the form of logistic regression is simple, the model is clear, the probability derivation behind it can withstand scrutiny, and the interpretability is very good.

\[
\hat{Y} = P(Y = 1|X)(X \in \mathbb{R}, 0 \leq \hat{Y} \leq 1) \tag{1}
\]

The logistic regression model function form is (2), where \( Z \) is a linear transformation, and the linear transformation can be closer to the predicted value of the true value \( Y \) after a certain transformation relationship. The transformation relationship here is the sigmoid function (3).

\[
\hat{Y} = \sigma(Z), \quad Z = w^TX + b(w, b \in \mathbb{R}) \tag{2}
\]
\[ \sigma(Z) = \frac{1}{1 - e^{-z}} \quad (3) \]

The calculation process first initializes \( w \) to a number that is randomly close to 0, \( b \) is equal to 0, and calculates the predicted output result \( Y \). The cost function is defined as the average of the loss function of \( m \) training samples, which measures the average error cost between the predicted result and the real result. The goal of optimization is to minimize the cost function (4).

\[
J(w, b) = \frac{1}{m} \sum_{i=1}^{m} L\left(y^{(i)}, \hat{y}^{(i)}\right) = -\frac{1}{m} \sum_{i=1}^{m} \left[y^{(i)} \log \hat{y}^{(i)} + (1 - y^{(i)}) \log (1 - \hat{y}^{(i)})\right] 
\quad (4)
\]

Use the gradient descent method to find the weight \( w \) and the offset \( b \) when the cost function is minimized. The gradient, the slope of the current point, specifies the direction of movement. The gradient descent method is to find the minimum value, so it moves in the negative direction of the gradient. Update \( w \) and \( b \) to be close to the lowest point of the cost function \( J \) curve, where \( \alpha \) is learning-rate representing the moving step length, which can be adjusted according to the model to obtain the optimal result. Within the number of iterations, repeat the above calculation steps until the optimal parameters \( w, b \) are obtained so that the cost function \( J(w, b) \) has a minimum value.

The function of the linear model is expressed as (5) and the loss function is expressed as (6), where \( n \) is the characteristic number.

\[
h(x) = \sum_{i=0}^{n} \theta_i x_i = \mathbf{\theta}^T \mathbf{x} 
\quad (5)
\]

\[
J(\theta) = \frac{1}{2} \sum_{i=1}^{m} \left(h_{\theta}(x^{(i)}) - y^{(i)}\right)^2 
\quad (6)
\]

In order to prevent over-fitting problems in the linear regression process, empirical risk minimization and structural risk minimization strategies are needed. The empirical risk minimization is to solve the optimization problem, such as solving the loss function minimization problem in linear regression, which is defined as (7). The structural risk minimization is equivalent to regularization, and a regularization term or penalty term representing the complexity of the model is added to the empirical risk. In the case of determining the loss function and training set data, it is defined as (8), where \( \lambda \) is called the regularization parameter.

\[
R_{\text{emp}}(f) = \frac{1}{N} \sum_{i=1}^{N} L\left(y_i, f\left(x_i\right)\right) 
\quad (7)
\]

\[
R_{\text{svm}}(f) = \frac{1}{N} \sum_{i=1}^{N} L\left(y_i, f\left(x_i\right)\right) + \lambda J(f) 
\quad (8)
\]

Lasso regression and Ridge regression both achieve the purpose by introducing regularization term into the loss function. Lasso regression introduces L1 norm penalty term, the sum of absolute values of each element of vector, and the loss function is (9). Ridge regression introduces the sum of squares of each element of L2 norm penalty term, and then calculates the square loss function as (10).

\[
R_{\text{svm}} = J(\theta) + \frac{\lambda}{2} \sum_{j=0}^{2} \theta_j^2 
\quad (9)
\]

\[
R_{\text{svm}} = J(\theta) + \lambda \sum_{j=0}^{2} \mid \theta_j \mid 
\quad (10)
\]
3.2 LightGBM

LightGBM is a framework that implements the GBDT algorithm, which solves the problem that GBDT cannot choose between data size and time consumption when processing data, and supports efficient parallel training [13-15]. The negative gradient of the loss function is used as the approximate residual value of the current decision tree to fit the new decision tree.

LightGBM uses the histogram algorithm, which occupies lower memory and has lower data separation complexity. The idea of the histogram algorithm is to convert continuous floating-point data into bin data, determine how many bucket bins are needed for each feature, and then divide them equally, update the sample data belonging to the bucket to the bin value, and finally represent it with a histogram. LightGBM discretizes continuous floating-point features into k discrete values, constructs a Histogram with a width of k, and then traverses the training data to count the cumulative statistics of each discrete value in the histogram. When performing feature selection, it is only necessary to traverse to find the optimal segmentation point according to the discrete value of the histogram. The most obvious advantage of using the histogram algorithm is the reduction in memory consumption, and the computational cost is also greatly reduced.

Boosting tree is an optimization process that uses additive model and forward distribution algorithm to achieve learning. It has some efficient implementations, such as XGBoost, pGBRT, GBDT (Gradient Boosting Decision Tree), etc. Their common shortcoming is that all samples need to be scanned when calculating information gain. When faced with a large amount of data or a high feature dimension, their efficiency and scalability are difficult to satisfy. The direct way to solve this problem is to reduce the amount of features and data without affecting accuracy. LightGBM is a good solution to these problems, mainly including Gradient-based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB) two algorithms.

The method used by GOSS is to reduce samples, which excludes a large part of the samples with small gradients, and uses the residuary samples to calculate the information gain, it is an algorithm that balances data reduction and accuracy. The gradient of the data is related to the specific value of each data instance, so the gradient of each data is different, and the information gain is more affected by the instance with a large gradient. Hence, when sampling, samples with large gradients should be kept as much as possible, and samples with small gradients should be randomly removed. Facts have proved that under the same sampling rate, especially when the information gain range is large, this measure can obtain more accurate results than random sampling. EFB starts from the perspective of feature reduction. High-dimensional data is often sparse. In a particularly sparse feature space, many features are almost mutually exclusive, and many features will not be non-zero at the same time. You can choose to bundle the two features. Replace it with a composite feature, then reduce the bundling problem to the graph coloring problem, and obtain an approximate solution through the greedy algorithm. Through EFB, many exclusive features are bundled into fewer dense features, greatly reducing the number of features, and improving training speed. The GBDT algorithm that combines GOSS and EFB is LightGBM. The algorithm principles of GOSS and EFB, the former reduces the number of samples, the latter reduces the number of features, combined with the principle of histogram algorithm to reduce the complexity of finding the optimal split point from three perspectives, and the tree growth strategy Optimized, these methods have improved the training speed of LightGBM.
4. Results and Discussion

In the binary classification problem, the sample has two categories: Negative and Positive. Then there are four combinations of model prediction results and real labels: TP (True Positive), FP (False Positive), FN (False Negative), TN (True Negative). Respectively means: the actual is a positive sample, and the prediction is a positive sample; the actual is a negative sample, and the prediction is a positive sample; the actual is a positive sample, and the prediction is a negative sample; the actual is a negative sample, and the prediction is a negative sample. Accuracy is the ratio of the number of correctly predicted samples to the total number of predicted samples, regardless of whether the predicted samples are positive or negative. Precision refers to the ratio of the number of correctly predicted positive samples to the number of all predicted positive samples, that is, how many of the samples predicted to be positive samples are truly positive samples. Precision only focuses on the part predicted to be positive samples, while accuracy considers all samples. Recall is the ratio of the number of correctly predicted positive samples to the total number of true positive samples, that is, how many positive samples can be correctly found from these samples. F1-score is equivalent to the harmonic average of precision and recall. If any value of recall and precision decreases, F1-score will decrease, and vice versa.

Select the above-mentioned encrypted text data to test the classification algorithm implemented by the system, and compare and analyze the various indicators and results of the algorithm through experiments. Most of the texts in the corpus are collected from encrypted newspapers and related media, and they are all related press releases. The prior classification of news releases is done by experts in various fields, and they are divided into 20 categories, which are first divided into social, economic, political, etc. according to their experience. We must first select the training set and the test set. Before that, the corpus that these experts have classified in advance is divided into 20 parts, and then one of them is selected as the open test set, and the remaining 19 parts are used as Training set and closed test set. Each copy can become an open test set, run the classification algorithm, and calculate the average value of the 20 classification operations. The experimental results are shown in Table 1. The overall result shows that the LightGBM effect is better.

| Method       | Precision | Accuracy | Recall | F1 score |
|--------------|-----------|----------|--------|----------|
| Logistic regression | 0.6       | 0.65     | 0.66   | 0.63     |
| LightGBM     | 0.7       | 0.73     | 0.72   | 0.73     |
5. Conclusion
This article mainly studies a method that can directly extract and classify encrypted text. The sample obtained is 9589, including the encrypted information and the result of classification. The number of categories to be classified is 20. Then the information is classified under the state of encryption, and keywords are extracted and classified based on the logistic regression and the LightGBM model. The model evaluation results are compared with precision, accuracy, recall and F1 score, showing that LightGBM is more effective.

References
[1] Aslett L J M, Esperança P M, Holmes C C. A review of homomorphic encryption and software tools for encrypted statistical machine learning[J]. arXiv preprint arXiv:1508.06574, 2015.
[2] Takabi H, Hesamifard E, Ghasemi M. Privacy preserving multi-party machine learning with homomorphic encryption[C]//29th Annual Conference on Neural Information Processing Systems (NIPS). 2016.
[3] Alves T, Das R, Morris T. Embedding encryption and machine learning intrusion prevention systems on programmable logic controllers[J]. IEEE Embedded Systems Letters, 2018, 10(3): 99-102.
[4] Khan K U, Qamar U. Improved Single-Label Text Categorization by Instance Filtration[C]// International Conference on Complex. IEEE, 2015.
[5] Keyvanrad M A , Homayounpour M M . A brief survey on deep belief networks and introducing a new object oriented toolbox (DeeBNet)[J]. Eprint Arxiv, 2014.
[6] I. Dilrukshi, K. De Zoysa, A. Caldera. Twitter news classification using SVM[C]// International Conference on Computer Science & Education. IEEE, 2013.
[7] Sun X, Zhang P, Liu J K, et al. Private machine learning classification based on fully homomorphic encryption[J]. IEEE Transactions on Emerging Topics in Computing, 2018.
[8] Vinayakumar R, Soman K P, Poomachandran P. Deep encrypted text categorization[C]//2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI). IEEE, 2017: 364-370.
[9] Sun X, Zhang P, Liu J K, et al. Private machine learning classification based on fully homomorphic encryption[J]. IEEE Transactions on Emerging Topics in Computing, 2018.
[10] Shan W, Zhang S, He Y. Machine learning based side-channel-attack countermeasure with hamming-distance redistribution and its application on advanced encryption standard[J]. Electronics Letters, 2017, 53(14): 926-928.
[11] Shan W, Zhang S, He Y. Machine learning based side-channel-attack countermeasure with hamming-distance redistribution and its application on advanced encryption standard[J]. Electronics Letters, 2017, 53(14): 926-928.
[12] Kim A, Song Y, Kim M, et al. Logistic regression model training based on the approximate homomorphic encryption[J]. BMC medical genomics, 2018, 11(4): 83.
[13] Aono Y, Hayashi T, Wang L, et al. Privacy-preserving deep learning via additively homomorphic encryption[J]. IEEE Transactions on Information Forensics and Security, 2017, 13(5): 1333-1345.
[14] Tanaka M. Learnable image encryption[C]//2018 IEEE International Conference on Consumer Electronics-Taiwan (ICCE-TW). IEEE, 2018: 1-2.
[15] Chialva D, Dooms A. Conditionals in homomorphic encryption and machine learning applications[J]. arXiv preprint arXiv:1810.12380, 2018.