Classification of Encrypted Text Based on Artificial Intelligence

Bo Zhang
Department of Computer Science, Illinois Institute of Technology, USA

Abstract. With the development of internet, information security is essential. The traditional decryption method only works for single encryption method. New models based on Artificial Intelligence to decrypt different kinds of encryption methods. Data features were extracted by tf-idf algorithm and applied into different models. The primary models to decrypt message without knowing whether encryption methods are logistic regression, lightGBM, and ensemble algorithm to decrypt encryption data. Comparing the performance in the models based on logistic regression and lightGBM algorithm respectively, we concluded that both the accuracy and F1 score of logistic algorithm are better than those of lightGBM algorithm. We also made improvement on the model by using ensemble algorithm. A more preferable performance including the accuracy and F1 score has achieved with multiple algorithm. Therefore, ensemble algorithm is more suitable for building general model to decrypting different kinds of encryption methods.

1. Introduction
With various usages of cryptology in society, researches on decryption and encryption methods became more and more important. Cryptology is used for securing digit signature, email, and even civil infrastructure [1-3]. The only way to determine the reliability of encryption methods is to test whether exists a short cut to decrypt. However, the tradition decryption is based on knowing encryption methods and each decryption method only works for one encryption methods [4-5]. Tradition methods cannot deal with a group of data with mixing encryption methods. Machine learning methods become more and more popular in those years and it is possible to use in decryption areas. In others fields, such as Go playing, natural language processing, and facial recognition, machine learning methods have proved its advantage [6-8]. The success in others fields motivates us to use machine learning methods to do research as decryption methods. We study performance of supervised model and unsupervised model to decryption data in order to find the best way to solve this problem.

Ronald L. Rivest proposed directions for future cross-fertilization and gave the relationship between cryptography and Artificial Intelligence [9]. Thore et al. used Linear Means (LM) Classifier and Fisher’s Linear Discriminant (FLD) Classifier to encrypt Wisconsin Breast Cancer Data set [10]. Probably approximately correct learning was used to decrypt message [11].

Thore Graepel showed levelled homomorphic encryption scheme and a machine learning algorithm on encrypted data [12]. A GA-Based Approach was adopted to reduce the cost of hide sensitive high utility itemsets [13]. Machine learning was used to analyze and encrypt data. It allowed arbitrary computations and the encrypted result to be better than general methods [14].

The primary purpose of this paper is to decrypt message without knowing encryption methods, using logistic regression, lightGBM, and ensemble algorithm to decrypt encryption data. Comparing the performance in different model, we will get the best general model. This paper contains 6 Sections. The first section is a general introduction to this program. Section 2 describes the process of data
research and Section 3 illustrates logistic regression and lightGBM algorithms in building models respectively. Accordingly, Section 4 shows the result of these two methods. In Section 5, a model has been built based on ensemble methods. Finally, the paper makes a conclusion of these algorithms when building models in Section 6.

2. Data Research
The dataset was originally collected by Ken Lang, containing evenly 20 Newsgroup. All the data are almost equally divided into 20 different newsgroups. Kaggle label all the encryption data into 4 difficult level, from 1 to 4. Level 1 means using the easiest encryption methods and level 4 means using the most difficult encryption methods. Decryption difficulty increases with the increment of level. The purpose of decryption is that using the ciphertext to predict the type of newsgroup.

The 20 newsgroups dataset are newspaper collections from around 20000 newsgroups documents. The topic of 20 newsgroups covers most of the topics such as computers, politics, religion and so on. One document only belongs to one groups. The training dataset is 39052 observations * 3 attributes and the 3 attributes are difficulty, ciphertext, and target. For the encryption dataset, the difficult level is equally randomly distributed. The following graph shows the distribution of 4 difficult levels.

![Distribution of 4 Difficulty Levels](image)

The next graph is the distribution of target columns. We can see that the target number are equally disturbed. Because of evenly distribution, precision became an important index to show the accuracy of our model.
Based on the property of encrypted text, we believe machine learning is a good approach to solve this problem, because ciphertext might have certain similar patterns [15-20]. Under this situation, logistic regression and lightgbm algorithm make good models with classification problems. Also, we use TF-IDF, term frequency–inverse document frequency method, to clean the encryption data in order to reduce noise and make a dictionary to store some particular patterns. After building models, we will use accuracy, precision and recall as our benchmark comparing our model to traditional decryption methods.

2.1 "TfidfVectorizer" Function
"TfidfVectorizer" function is used in python to extract similar patterns. This function convert document to TF-IDF feature. Tf feature calculates the frequency of the words. Idf reduces the weights of words such as “as”, “of”, “that” because those words are meaningless [5]. TF-IDF feature is equals to tf matrix times idf matrix. TF-IDF feature was used as important attribute in regression and lightgbm algorithm.

\[
\text{tf}(t,d) = 0.5 + 0.5 \times \frac{f_{t,d}}{\max \{ f_{t,d'} : t' \in d \} }
\]

\[
\text{idf}(t,D) = \log \frac{N}{| \{ d \in D : t \in d \} |}
\]

\[
\text{tfidf}(t,d,D) = \text{tf}(t,d) \times \text{idf}(t,D)
\]

2.2 Logistic Regression
Logistic regression is a widely used algorithm in classification problem, such as cancer prediction. Logistic regression is a supervised model which means the model will best fit the train set. The logistic regression model is built by minimizing the cost function. By minimizing the cost function, the model will have less sum of residual. Logistic regression can easily implement, fastly build up a model, and directly explains the weights of different features. Unlike deep leaning algorithm, logistic regression can fully demonstrate the relationship between each feature. Also, logistic regression can efficiently update model with new data. Using L1 and L2 normalization, logistic regression can solve multicollinearity phenomenon in data set.
The purpose of L1 and L2 lost function is to normalized dataset, preventing overfitting. The logistic regression model was built by minimizing the residual error. However, there are some disadvantages in logistic regression. It only limits to certain dataset comparing with decision tree model. The performance of logistic regression is poor dealing with large scale of data features or multiple dimension. The assumption of logistic regression is that input variables have linear relation with results.

2.3 Lightgbm
LightGBM model was original from XGBoost model. It adapts the advantage of XGBoost model and reduce training time comparing with XGBoost and it contains L1 and L2 lost function inside the model [9]. Lightgbm is derived from gradient boosting decided tree (GBDT). GDBT returns result by multiple week classification models [21]. The GDBT model has high accuracy rate and without overfit or underfit. Lightgbm model contains the advantage and ameliorates the defects of GDBT model. The improvement of Lightgbm model are faster building rate, least memory use, better results, and fastly dealing with large scales of data [22]. In Lightgbm converts data features into a histogram and uses the histogram to find the best diving points. Different from “level-wise” strategy in GDBT model, Lightgbm uses “leaf-wise” strategy. “leaf-wise” strategy will split value leaf and discard useless leaf. Dealing with large scale of data, lightgbm algorithm makes an appreval on parallel computing [23]. Traditional algorithm splits data evenly and distributes data into different threads. Each thread calculates its own data and find the best splits points. Accumulating all the local best splits points, traditional algorithm will return the global splits point. However, this way will waste plenty of resource on communication. In LightGBM algorithm, each thread contains all the data and uses data to calculate the best splits points.

2.4 Evaluation
Accuracy rate and F1 score were used to evaluate our model. Accuracy is percentage of correctness and F1 score represents harmonic mean of precision and recall. The range of accuracy and f1 score is from 0 to 1. Better value means better quality of the models.

\[
\text{Accuracy} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}
\]

\[
F1 \text{ Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}
\]

3. Result
Data features was extracted by TF-IDF algorithm. 80% of the data was randomly chosen as training dataset and the rest of the data was as testing dataset. Training dataset is used to build model and testing dataset is used to test the accuracy of the model. Accuracy rate and F1 score use as benchmark of the predicting results.

In this section, the following tables shows the results of different models. Table 1 is corresponding with logistic regression model, LightGBM model and ensemble model. In general, logistic regression
have the best overall accuracy rate. This is because ciphertext has high linear correlation with results. On the other hand, the accuracy rate and F1 score of the level 3 and level 4 are low. The linear correlation between attribute and results decreases when the level is high.

Table 1 Evaluation of Different Models.

|              | Logistic L1 | Logistic L2 | LightGBM |
|--------------|-------------|-------------|----------|
| Accuracy 1   | 0.83        | 0.82        | 0.74     |
| Accuracy 2   | 0.72        | 0.65        | 0.68     |
| Accuracy 3   | 0.63        | 0.53        | 0.43     |
| Accuracy 4   | 0.42        | 0.35        | 0.30     |
| F1 1         | 0.79        | 0.80        | 0.74     |
| F1 2         | 0.66        | 0.61        | 0.63     |
| F1 3         | 0.56        | 0.47        | 0.38     |
| F1 4         | 0.33        | 0.31        | 0.31     |

4. Ensemble
The accuracy of the result is highly depend on the model. Single model processing might not represent the full view of the data. Ensemble L1 model and lightgbm model will help to improve accuracy. The results of logistic regression will be a new input attribute in LightGBM model. The ensemble model utilizes the advantages from both model and prevents overfitting problem. Figure 3 fully explains the ensemble model.

Figure 3  Fusion of models

The ensemble model was created by the result of L1 model as a new input attribute and the original dataset. In the ensemble model, it can reduce overfitting problem. The accuracy rate and F1 score of the level 1 to 4 are low, which are better than other models stated above.

Table 2 Evaluation of Fusion Model.

|              | Ensemble |
|--------------|----------|
| Accuracy 1   | 0.85     |
| Accuracy 2   | 0.73     |
| Accuracy 3   | 0.63     |
| Accuracy 4   | 0.44     |
| F1 1         | 0.84     |
| F1 2         | 0.71     |
| F1 3         | 0.51     |
| F1 4         | 0.39     |

5. Conclusion
In conclusion, after cleaning data and extracting feature from data set, we used logistic regression, lightGBM and ensemble algorithm to build model. Comparing with accuracy rate and F1 score within different models, we suggested the result logistic regression model are the best. The reason could be that the distribution of data fits with logistic regression model.
We can make an improvement on feature extraction. Also, using principal component analysis method to reduce the dimension of data set would be a good choice because of many features and prevent overfitting. Multiple model ensemble will improve the accuracy of the model. In the future, we can use string match algorithm to compare with string and extract more value information.

References
[1] MACQ, B. M, QUISQUATER, et al. Cryptology for digital TV broadcasting[J]. Proc IEEE, 1995, 83(6):944-957.
[2] Nguyen P Q, Stern J. The Two Faces of Lattices in Cryptology[C]// Revised Papers from the International Conference on Cryptography & Lattices. 2001.
[3] Simmons G J. Contemporary cryptology-The science of information integrity[J]. Designs Codes & Cryptography, 1994, 4(2):193-195.
[4] Sahai A, Waters B. Fuzzy Identity-Based Encryption[C]// International Conference on Theory & Applications of Cryptographic Techniques. 2005.
[5] Needham R M. Using encryption for authentication in large networks of computers[M]. 1978.
[6] Goldberg D E. Genetic algorithms in search, optimization and machine learning[J]. 1989.
[7] Murphy K P. Machine learning : a probabilistic perspective[J]. Chance, 2012, 27(2):62-63.
[8] Blum A L, Langley P. Selection of relevant features and examples in machine learning[J]. Artificial Intelligence, 1997, 97(1–2):245-271.
[9] Rivest R L. Cryptography and machine learning.[C]// International Conference on the Theory & Applications of Cryptology: Advances in Cryptology. 1991.
[10] Wong W K, Manzur M, Chew B K. How rewarding is technical analysis? Evidence from Singapore stock market[J]. Applied Financial Economics, 2003, 13(7): 543-551.
[11] Ke G, Meng Q, Finley T, et al. Lightgbm: A highly efficient gradient boosting decision tree[C]/Advances in Neural Information Processing Systems. 2017: 3146-3154.
[12] Graepel T, Lauter K, Naehrig M. ML Confidential: Machine Learning on Encrypted Data[C]// International Conference on Information Security & Cryptology. 2012.
[13] Lin C W, Hong T P, Wong J W, et al. A GA-based approach to hide sensitive high utility itemsets.[J]. The Scientific World Journal, 2014, (2014-3-3), 2014, 2014(1):804629.
[14] Anderson B, Mcgrew D. Machine Learning for Encrypted Malware Traffic Classification: Accounting for Noisy Labels and Non-Stationarity[C]// the 23rd ACM SIGKDD International Conference. 2017.
[15] Kim M, Song Y, Shuang W, et al. Secure Logistic Regression Based on Homomorphic Encryption: Design and Evaluation[J]. Jmnr Medical Informatics, 2018, 6(2).
[16] Chouhan P, Richhariya V, Chouhan P, et al. A Survey: Analysis of Current Approaches in Anomaly Detection[J]. Foundation of Computer Science, 2015, 111(17):14-20.
[17] Graepel T, Lauter K, Naehrig M. ML Confidential: Machine Learning on Encrypted Data[M]// Information Security and Cryptology – ICISC 2012. 2012.
[18] Rajaraman, A.; Ullman, J.D. (2011). “Data Mining” (PDF). Mining of Massive Datasets. pp. 1–17. doi:10.1017/CBO9781139058452.002. ISBN 978-1-139-05845-2.
[19] Laskari E C, Meletiou G C, Tasoulis D K, et al. Data Mining and Cryptology[C]// International Conference on Computational Methods in Sciences & Engineering. 2003.
[20] Park H A, Dong H L, Zhan J, et al. Efficient keyword index search over encrypted documents of groups[C]// IEEE International Conference on Intelligence & Security Informatics. 2008.
[21] Kadiyala A, Kumar A. Applications of python to evaluate the performance of decision tree - based boosting algorithms[J]. Environmental Progress & Sustainable Energy, 2018, 37(2):618-623.
[22] Kai C, Yi Z, Li C, et al. A Learning-based Text Detection Method in Camera Images[J]. IEEE, 2011, 2:270 - 274.
[23] Li K, Jr E H W. Method and apparatus for text - independent speaker recognition[J]. Journal of the Acoustical Society of America, 1990, 88(5):2513-2513.