Filtering Network Spam Message using Approximated Logistic Regression

Yiping Zheng and F. Liu
Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education of China, Xidian University, Xi'an 710071, P.R. China

Abstract—The development of telecom network and Internet provides effective ways for communication. As an important way in communication, Short Messaging Service (SMS) via both telecom network and Internet has played an increasing important role in daily life. However, it usually suffers from spam SMS that causes misunderstanding and cheat. The highly varying content, network environment make the identification of spam message difficult. Although the previous methods to some extent can filter the spam messages, it usually fails to capture the semantic information because it simply relies on keywords. Thus, its accuracy is not satisfied enough. Also, their further applications to some difficult situations of spam SMS filtering are still limited by their shortcomings, i.e., their adaptation ability to network environment and their robustness to noise. Therefore, high efficiency spam SMS filtering method is of greatly important. In this paper, to overcome the shortcomings of previous methods for spam message filtering, we propose a new approach, linear discrimination based keyword selection with approximated logistic regression (KW-ALR). The proposed approach KW-ALR first extracts feature or keywords using linear discrimination analysis, and then trains spam recognition model based on approximated logistic regression over the extracted keywords. We evaluate the proposed approach KW-ALR over a standard data set SMS Spam Collection. The experimental result shows that our method KW-ALR for spam message filtering achieves higher accuracy over other methods.

Index Terms—Spam Message; Second Term; Third Term; Fourth Term; Fifth Term; Sixth Term

I. INTRODUCTION

With the constantly strengthening and improvement of the wireless communication service, SMS provides users a convenient messaging service as well a convenient channel for spreading spam information [1, 2, 3]. The existence of spam messages has brought a lot of troubles to users and operators, such as menace and harassment for users. And the operators would be misunderstood by the users due to users’ fault judgment of the message sources, which seriously affect the operator's reputation. In addition, if the balance of phone is insufficient, in this period of using time delay deduction system, the frequent and large number of message and the malicious arrearage will bring economic losses to operators [3, 6]. In addition, the high attention of media makes the related telecom regulator and the telecom operators taking great pressure [5, 6, 7, 8]. Guaranteeing the communication rights of each user and giving restriction and punishment accurately and timely for the malicious users is highly challenging, e.g., how to define the spam messages efficiently. Thus, whether blocking spam behavior becomes a public concern.

As an important way of communication, SMS has brought great convenience to daily communication. With the widespread of text messages, spam messages become new language pollution. The hidden spam messages and changeable content, text form and frequency make it difficult being. The methods relying on keywords or flow monitor cannot satisfy the requirement of spam filtering of the operators. Therefore, developing highly efficient recognition methods is significant for the recognition of spam messages.

Spam messages have some unique features, such as advertising, fraud, pornography, curse, etc. The details of spam messages can be described by the following properties. SMS identification is usually based on the characteristics of the text, needing to select the characteristic attributes, such as syntax and grammar, sentence patterns, semantics. [5] selects more than two punctuation properties from the perspective of syntax and grammar, and select text length attribute from the perspective of the length of the text sentence structure; select keywords attributes, including keywords attribute, sending information or replying information attribute, telephone number or address information attribute from the perspective of semantics attributes. It sets weights these attributes to represent their contribution in determining spam message [8]. Therefore, extract the characteristic factors of text message from the attribute level can filter the spam messages. On the other hand, spam messages recognition is a nonlinear problem. In the previous approaches, most of the solutions are to set the blacklist. But it will produce some errors by using the linear classification method when filter the content of the text. And, to handle a larger number of spam messages, the computation cost is expensive and the processing speed is slow.

Specifically, the previous approaches mainly analyzed the text, using the artificial neural network, support vector machine (SVM), naïve Bayes algorithm to predict and filter spam messages. Due to the great amount concurrence of SMS in the actual processing of these algorithms, it requires that the system is highly efficient.
Among them, the artificial neural network method can deal with fuzzy information classification, but its classification accuracy is not satisfied enough.

Support vector machine defines each word as a dimension of the space. When the document is converted to a vector of the text space, the value of the vector of each dimension should be expressed as the weight of corresponding word in the document and then define the document classification by comparing the text distance. But when calculating the weight of words in the text, due to the short length of message content, word repeating almost never happen. Bayesian method calculates the posterior probability to classify the text based on word frequency. But when the message is expressed as vector, the lack of phenomenon is serious, which difficult to distinguish on distance. When using these methods to conduct the recognition on spam messages, the computational complexity is high, with spam messages hooking on the system. For the users, the influence brought by the SMS system performance and the frequency and range of text hooking would be reduced, which is not accurate enough for the spam messages.

Here we make a summary of our proposed algorithm KW-ALR. It mainly contains four steps. (1) Remove the noise in SMS so as to minimize the ambiguity and highlight the useful information. (2) Extract features from the processed data using linear discrimination analysis, with the data distribution considered. (3) Train approximated logistic regression model for spam message identification, with the similarity measure over the keyword feature space embedded. (4) The identification procedure is conducted on the basis of learnt KW-ALR where the keyword features for test data are extracted in the same way of training data.

By means of comparing with previous algorithms for spam message identification, it highlights the main contributions of this paper in three aspects. (1) We propose to model the distribution of SMS and extract keyword features based on the distribution, which provides strong adaption ability. (2) We employ approximated logistic regression as the model for identification, which exploits its ability in dealing with nonlinear problem and benefits from its optimum solution. (3) The experimental results show that the proposed KW-ALR outperforms other related methods for spam message identification, on different settings and criterions, which verifies its advantages.

The remaining part of the paper is organized as follows. Part 2 reports our developed method KW-ALR. We verify our proposed approach KW-ALR and other methods in Part 3. Section 4 draws a conclusion.

II. OUR PROPOSED SCHEMA

The previous algorithms suffer from the defects described in above section. To overcome their limitations for spam message identification, we propose an algorithm, so called KW-ALR, on the base of the linear discrimination analysis and approximated logistic regression. For readability, we illustrate our developed KW-ALR in Figure 1. As shown in Figure 1, it mainly is composed of three procedures. Firstly, collect and process data. Secondly, feature extraction using the approach in the above section. Thirdly, train the model and perform test.

A. The Spam SMS Dataset

The SMS Spam Collection used in our experiments was collected by Tiago A. Almeida of Federal University of Sao Carlos. It was employed to identify the text of spam messages, which is a challenging and time consuming task, involving the scanning of hundreds of web pages. This corpus has been collected from free or free for research sources at Internet: a collection of 425 SMS spam messages was manually extracted from the Grumble text Web site. This is a UK forum in which cell phone users make public claims about SMS spam messages, most of them without showing the very spam message received. The spam message identification is a very hard and time consuming task, and it involved carefully scanning hundreds of web pages.

A subset of 3375 SMS are stochastically chosen from NUS SMS Corpus (NSC), which is a dataset of about 10,000 legitimate messages collected for research at the Department of Computer Science at National University of Singapore. The messages largely originate from Singaporeans and mostly from students attending the University. These messages were collected from volunteers who were made aware that their contributions were going to be made publicly available. A list of 450 SMS ham messages collected from Caroline Tag's PhD. Finally, we incorporated the SMS Spam Corpus v.0.1 Big. It has 1002 SMS ham messages and 322 spam messages and it is public available. The database includes have 5574 examples, these attributes describes as spam and ham. The classes description is presented in Table I.

<table>
<thead>
<tr>
<th>Categories</th>
<th>Number of examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spam</td>
<td>747</td>
</tr>
<tr>
<td>Ham</td>
<td>4827</td>
</tr>
<tr>
<td>Total</td>
<td>5574</td>
</tr>
</tbody>
</table>

B. Keyword Selection Using LDA

Linear discriminant analysis (LDA) [8] is widely used in pattern recognition and machine learning. It is usually used to determine a linear combination of features. The resulting combination can be used for dimensionality reduction for later identification. The equivalent technique is discriminant correspondence analysis when dealing with categorical independent variables. Considering a configured of observations $x$, for each datum of an object is associated with known class label.
y. Then the identification problem is to determine a predictor for the class y of arbitrary example of the same distribution given only an observation x. LDA assumes that both the conditional probability density functions $p(x | y = 0)$ and $p(x | y = 1)$ are normally distributed with parameters $(\mu_0, \Sigma_{y=0})$ and $(\mu_1, \Sigma_{y=1})$ respectively. Under this assumption, the Bayes optimal solution is to predict points as being from the second class if the log likelihood ratios is below some threshold $T$, so that:

$$
(x - \mu_b)^T \Sigma_{y=0}^{-1} (x - \mu_b) + \ln |\Sigma_{y=0}| - (x - \mu_i)^T \Sigma_{y=1}^{-1} (x - \mu_i) - \ln |\Sigma_{y=1}| < T
$$

We assume that $\Sigma_{y=0} = \Sigma_{y=1} = \Sigma$ and the covariance has full rank. In this case, the above decision rule becomes a threshold $w^T x > c$ for some threshold $c$. It implies that the criterion of an input $x$ being in a category $y$ is a function of this linear combination of observations.

Under the situations that there are more than two categories, we can extend the analysis make use of in the derivation of the Fisher discriminant to find a subspace that appears to include all of the category variability. This generalization is result from [9]. Suppose that each of $C$ categories has a mean $\mu_i$ and the same covariance $\Sigma$.

Then we can define the inter-class variance by the instance covariance of the category demonstrates

$$
\sum_n = \frac{1}{n} \sum_{i=1}^{n} (\mu_i - \mu)(\mu_i - \mu)^T
$$

where $\mu$ is the mean of all class. The category separation in a direction $w$ in this situation will be given by means of,

$$
S = \frac{w^T \sum_n w}{w^T \Sigma w}
$$

It implies that, when $w$ is an eigenvector of $\Sigma^{-1} \Sigma_n$, the separation will be equal to the corresponding eigenvalue. If $\Sigma^{-1} \Sigma_n$ is diagonalizable, the variability in features will be included in the subspace spanned by the eigenvectors corresponding to the $C-1$ largest eigenvalues. We take the advantage of these eigenvectors in feature reduction. The eigenvectors corresponding to the small eigenvalues will tend to be very sensitive to the exact choice of training sample, and it is often necessary to employ regularization, that described in the next section.

LDA can be used to select key words through the following procedures. First, train LDA using the training set. The direction $w$ weights the importance of words in spam SMS identification. We select those words with a large absolute weight as the keywords for spam SMS identification.

C. Approximated Logistic Regression for Recognition

Section 2.2 selects keywords from a large number of words. The words then will be used to construct the representation of SMS. That is, we count the frequency of keyword appeared in the SMS, and use the histogram of word occurrence as feature to represent SMS. The feature is then delivered to the following classifier for identification.

Logistic Regression can predict both real and binary responses, where the output posterior probabilities can be processed expediently and sent to other systems. It attempts to simulate the category label’s conditional probability based on its observation:

$$
p(y | x) = \frac{1}{1+\exp(-y(w^T x + b))}
$$

where $x = (x_1, \ldots, x_m)^T$ is the data vector; $m$ is the number of features; $y \in \{+1,-1\}$ is the binary-valued class label; $w = (w_1, \ldots, w_m)^T$ is the weight vector; $b$ is decision intercept. The weight can be calculated using:

$$
\hat{w} = \arg \min_w \frac{1}{n} \sum_{i=1}^{n} \log [1+\exp(-y_i (w^T x_i + b))] + \lambda \sum_{i=1}^{m} w_i^2
$$

Here we first construct a series of optimization problems whose solutions converge to the solution of SVM. Therefore, SVM can be solved using simple unconstrained optimization techniques. Then we put forward our simple ALR-CG approach [10] which employs CG as its inner loop.

So as to simplify the formulation, we adopt the augmented weight vector $w = (b, w_1, \ldots, w_m)^T$ and the augmented example vector $x = (1, x_1, \ldots, x_m)^T$ from right now unless otherwise specified. To keep the SVM optimization problem unchanged, its form becomes,

$$
\hat{w} = \arg \min_w \left\{ \frac{1}{n} \sum_{i=1}^{n} \max\{0,1-y_i w^T x_i\} + \lambda \sum_{i=1}^{m} w_i^2 \right\}
$$

The intercept $\omega_0 = b$ is not in the regularization term.

We also need not to penalize the intercept $\omega_0$ in the regularized LR to approximate SVM:

$$
\hat{w} = \arg \min_w \left\{ \frac{1}{n} \sum_{i=1}^{n} \log [1+\exp(-y_i w^T x_i)] + \lambda \sum_{i=1}^{m} w_i^2 \right\}
$$

It is known that the loss functions play an important role in SVM and LR. The SVM loss function can be approximated by means of the loss of the following approximated LR [12]:

$$
g_{\gamma}(x, y, w) = \frac{1}{\gamma} \ln [1+\exp(-\gamma (yw^T x - 1))] \tag{1}
$$

If we can approximate the SVM loss function,

$$g_{\text{svm}}(x, y, w) = \max\{0,1-yw^T x\}
$$

The problem can be resolved using simple unconstrained optimization techniques.

It can be proven that, we can solve SVM by solving the problems a sequence of sub-optimization. Here we solve each sub-optimization problem using the Conjugate Gradient (CG). For finding large-scale nonlinear optimization problems, CG [13] is one of the most
popular approaches. More importantly, [14] compared it with other approaches in fitting LR, and solved which it is more efficient than other algorithms.

III. EXPERIMENTS

This section will empirically assess our developed KW-ALR algorithm for spam message identification. We will introduce the database, verification standard and the experimental results sequentially. The experimental step of the proposed KW-ALR algorithm, in both mathematics and graphical form, can be determined in the above part. KW-ALR mainly contains three steps: (1) collect data according to experiment design; (2) feature extraction via approach in Section 2; (3) train the model and evaluate its performance [14].

A. Verification Standard

To verify the superiority of the proposed algorithm KW-ALR for spam message identification, we compare it with KW-LR, Ad boost and SVM, and utilize the evaluation standards, recognition accuracy and precision, for verification towards spam message identification. The definitions of these standards for spam message identification are defined as follows:

\[
\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}
\]

\[
\text{Precision} = \frac{TP}{TP + FP}
\]

FP is the short of false positive, representing the unexpected predicted positive instances of SMS; FN is the short of false negative, denoting the unexpected predicted negative datum of SMS; TP is the short of true positive and TN is the short of true negative. These assessment criterions are defined under the situation of two class problem of spam message identification, and can be utilized for multiple category problem of spam message identification.

B. Experimental Results

In this experiment, we validate the advantage of our proposed KW-ALR algorithm in spam message identification, through the comparison experiments with KW-LR and Ad boost etc. Two popular criterion, accuracy and precision, are adopted for assessment. Identification Accuracy and Precision are two typical and popular measures for the performance of the identification model. The dataset is SMS Spam Collection. The experimental steps are shown in above section. Our method KW-ALR is learnt by the approach in above section, where some parameters of KW-ALR are set to defaults. The test of spam message identification is repeatedly performed for several rounds over stochastically partition dataset.

We conduct the test experiment of spam message identification by taking advantage of our developed algorithm over 10 stochastically partition training sets and test configures, and report the identification performance in Table II and Figure 2. As shown in Table II and Figure 2, the accuracy of our developed approach, for varying percentage of training examples, varying experimental configurations and distinct verification criterions, exhibits competitive performance, i.e., higher than others about 82.85%-79.29%. The performance ranges over accuracy and precision are respectively 70.33%-89.40% and 69.88%-89.33%. These results mean that our proposed method is robust to the percentage of training instances of SMS. The reasons accounting for the above results are mainly from the following three aspects. First, the KW-ALR method has the ability to map the linearly inseparable data to the separable feature in high dimensional space via the feature mapping based on linear discrimination analysis, which makes the classification problem easy. Second, the parameter selection method selects the parameters of KW-ALR based on the distribution of SMS data, which makes the KW-ALR having better adaptability. Third, the frame of our proposed method KW-ALR contains a group of comprehensive steps which sequentially maximize the identification ability [15].

<table>
<thead>
<tr>
<th>Training samples</th>
<th>Method (Accuracy)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>KW-ALR</td>
</tr>
<tr>
<td>30%</td>
<td>70.33</td>
</tr>
<tr>
<td>40%</td>
<td>76.61</td>
</tr>
<tr>
<td>50%</td>
<td>82.85</td>
</tr>
<tr>
<td>60%</td>
<td>87.07</td>
</tr>
<tr>
<td>70%</td>
<td>89.40</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Training instances</th>
<th>Method (Precision)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>KW-ALR</td>
</tr>
<tr>
<td>30%</td>
<td>69.88</td>
</tr>
<tr>
<td>40%</td>
<td>78.37</td>
</tr>
<tr>
<td>50%</td>
<td>83.55</td>
</tr>
<tr>
<td>60%</td>
<td>86.88</td>
</tr>
<tr>
<td>70%</td>
<td>89.33</td>
</tr>
</tbody>
</table>

Figure 2. The identification results of spam message identification utilizing KW-LR and KW-ALR.

Here, we target to verify the advantage of our developed KW-ALR for spam message identification, and the capability of the feature extraction algorithm on the basis of linear discrimination analysis (LDA). The experiments are conducted over SMS Spam Collection, comparing with KW-LR, Adboost and SVM. The data in our experiment were collected by Tiago A. Almeida. The dataset has 5574 samples which are divided into two parts, training data set with 5000 samples and forecasting data set with 574 samples. The used evaluation criterions are accuracy and precision where identification accuracy and precision are two typical and popular measures for the
performance of the identification model. The experimental procedures are presented in the model part, where the KW-ALR is determined by the criterion algorithm, with the parameters configured by authors.

We target to evaluate the advantage of our developed KW-ALR for spam message identification, through extensively comparing our proposed KW-ALR with three state-of-the-art algorithms, which is, KW-LR, Ad boost and SVM. In addition, to assess the robustness of our proposed KW-ALR against the number of training SMS instances, we also vary the number of training samples. The experimental results are summarized in Table III. As shown in Table III, the proposed algorithm KW-ALR consistently outperforms all three compared approaches, about 19.07% on accuracy and 19.45% on precision. Moreover, our developed KW-ALR exhibits strong robustness against the percentage of training instances from 30% to 70%. These results indicate that, our KW-ALR can be straightforwardly applied to a number of applications. We here provide some explanations. (1) Comparing with the previous methods (Adaboost, KW-Lr and SVM), our KW-ALR based on linear discrimination analysis and approximated logistic regression can deal with the situation that has a large number of samples in high dimensional space. (2) The optimization method of our proposed KW-ALR is robust against the initial parameters, and thus could reach a stable solution, compared with previous methods like KW-LR. (3) The processing step of SMS data is able to remove noise while keeping useful information effectively. Moreover, the element steps of KW-ALR cooperate with each other.

In this experiment, we seek to validate the ability and robustness of KW-ALR algorithm for spam message identification, in comparison with related approaches. The SMS Spam Collection is collected from Grumble text Web site and is stochastically split to training set and test set. The dataset in our experiment were collected by Tiago A. Almeida. The dataset has 5574 samples which consist of four steps. (1) Remove the noise in SMS so as to minimize the ambiguity and highlight the useful information. (2) We extract the feature on the processed data, through linear discrimination analysis that is based on the distribution. (3) Train the approximated logistic regression model for identification, with the similarity measure over the keyword feature space embedded. (4) We perform identification for spam message identification using the learned KW-ALR and extracted keyword features for test data. Compared with previous algorithms for spam message identification, the main contributions of this work are three aspects. (1) We propose to extract keyword features based on the distribution of SMS data, which can take the variance of SMS data into account. (2) Approximated logistic regression (ALR) that used for identification is able to fully exploit the informative components of extracted keyword features, be means of its feature selection ability. (3) The experimental results show that KW-ALR outperforms other related methods for spam message identification, on different configurations and criterions, which verifies its advantages. As the future paper,

### Table III. Identification Performance Comparison for Spam Message Identification Over Four Approaches

<table>
<thead>
<tr>
<th>Training example</th>
<th>KW-LR</th>
<th>Adaboost</th>
<th>SVM</th>
<th>KW-ALR</th>
</tr>
</thead>
<tbody>
<tr>
<td>30%</td>
<td>66.96</td>
<td>65.93</td>
<td>68.37</td>
<td>70.33</td>
</tr>
<tr>
<td>40%</td>
<td>71.45</td>
<td>75.77</td>
<td>72.37</td>
<td>76.61</td>
</tr>
<tr>
<td>50%</td>
<td>79.29</td>
<td>78.67</td>
<td>76.88</td>
<td>82.85</td>
</tr>
<tr>
<td>60%</td>
<td>84.34</td>
<td>79.82</td>
<td>78.24</td>
<td>87.07</td>
</tr>
<tr>
<td>70%</td>
<td>84.74</td>
<td>82.84</td>
<td>82.01</td>
<td>89.40</td>
</tr>
</tbody>
</table>

### Table IV. The Experimental Results of SVM and KW-ALR

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Verification standard</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM</td>
<td>74.79</td>
</tr>
<tr>
<td>KW-LR</td>
<td>83.09</td>
</tr>
<tr>
<td>SVM</td>
<td>76.59</td>
</tr>
<tr>
<td>KW-ALR</td>
<td>83.80</td>
</tr>
<tr>
<td>SVM</td>
<td>74.96</td>
</tr>
<tr>
<td>KW-ALR</td>
<td>83.45</td>
</tr>
<tr>
<td>SVM</td>
<td>76.98</td>
</tr>
<tr>
<td>KW-ALR</td>
<td>82.60</td>
</tr>
<tr>
<td>SVM</td>
<td>78.28</td>
</tr>
<tr>
<td>KW-ALR</td>
<td>84.37</td>
</tr>
</tbody>
</table>

### IV. Conclusions

As a summary, we propose an algorithm KW-ALR on the basis of linear discrimination analysis and approximated logistic regression. The approach mainly is composed of four steps. (1) Remove the noise in SMS so as to minimize the ambiguity and highlight the useful information. (2) We extract the feature on the processed data, through linear discrimination analysis that is based on the distribution. (3) Train the approximated logistic regression model for identification, with the similarity measure over the keyword feature space embedded. (4) We perform identification for spam message identification using the learned KW-ALR and extracted keyword features for test data. Compared with previous algorithms for spam message identification, the main contributions of this work are three aspects. (1) We propose to extract keyword features based on the distribution of SMS data, which can take the variance of SMS data into account. (2) Approximated logistic regression (ALR) that used for identification is able to fully exploit the informative components of extracted keyword features, be means of its feature selection ability. (3) The experimental results show that KW-ALR outperforms other related methods for spam message identification, on different configurations and criterions, which verifies its advantages. As the future paper,
KW-ALR may benefit from the two aspects: (1) more sophisticated optimization method will improve the computational efficiency of KW-ALR. (2) Our developed KW-ALR can be applied to other real world tasks.

V. REFERENCES


