Engineering Applications of Artificial Intelligence 75 (2018) 11–21

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

journal homepage: www.elsevier.com/locate/engappai

Email security level classification of imbalanced data using artificial neural network: The real case in a world-leading enterprise

Jen-Wei Huang [a,*], Chia-Wen Chiang [b], Jia-Wei Chang [c]

[a] Department of Electrical Engineering, National Cheng Kung University, 70101, Tainan, Taiwan ROC
[b] Institute of Computer and Communication Engineering, National Cheng Kung University, 70101, Tainan, Taiwan ROC
[c] Department of Computer Science and Information Engineering, National Taichung University of Science and Technology, 40401, Taichung, Taiwan ROC

ARTICLE INFO

Keywords:
E-mail
Classifier
Text mining
Artificial neural network

ABSTRACT

Email is far more convenient than traditional mail in the delivery of messages. However, it is susceptible to information leakage in business. This problem can be alleviated by classifying emails into different security levels using text mining and machine learning technology. In this research, we developed a scheme in which a neural network is used to extract information from emails to enable its transformation into a multidimensional vector. Email text data is processed using bi-gram to train the document vector, which then undergoes under-sampling to deal with the problem of data imbalance. Finally, the security label of emails is classified using an artificial neural network. The proposed system was evaluated in an actual corporate setting. The results show that the proposed feature extraction approach is more effective than existing methods for the representations of email data in true positive rates and F1-scores.

1. Introduction

Email is a common way to communicate with others nowadays, especially in a corporation and between companies. Employees usually receive and send lots of information via emails every day. However, the enormous amount of emails has brought some difficulties, such as spamming (Ismaila et al., 2015), privacy threat (Snchez and Batet, 2017) or information leakage. Especially, when employees send emails to people outside the company, they may accidentally or intentionally include some confidential information. If managers do not scrutinize the outbound emails, the sensitive information will be exposed to the public or the competitive company, which will result in significant loss of the corporate. However, it is a massive task for managers because the vast numbers of emails are required to be inspected within one day.

To address this urgent problem, this study proposes an effective system to assist in the security level classification of corporate emails using text mining and machine learning techniques. The proposed system can scrutinize lots of contents of emails and prevent sensitive information from leaking based on the emails security labels. In the corporation, all emails will be examined by the proposed email security classification system before being sent to other companies. If the classified security level does not match the label provided by email users, the email will be suspended from sending out and reported to the corresponding manager. In this way, the amount of email needed to be checked by managers can be reduced significantly.

However, there are practical issues needed to be solved: (1) Due to the privacy policy of the corporation (subject), we cannot use the meta-data of emails, including senders and receivers. The only available information to classify emails is text. Under the very limited information, Emails are required to be classified into different security levels according to the textual contents of the emails and the attachments. (2) The portion of sensitive email data in the real world is much less than those of usual email data, which results in significant differences between the amount of data in each class. Therefore, the proposed method must address the data imbalance problem. (3) A DNN model with excellent performance requires enormous amounts of training data. Dealing with such volumes of data imposes considerable difficulties concerning time and computational overhead. In our test case, there are more than 150 thousands of emails which are short and imbalanced.

To conquer three major problems mentioned above, we proposed the corresponding methods to them as follow: (1) We need to extract the meaningful textual features of the email body and the attachments. However, emails are usually short to have very limited information. Therefore, we improve the neural network based paragraph vector (Le and Mikolov, 2014) to represent processed emails. The traditional paragraph vector disregards the order of individual words in the text.

* Corresponding author.
E-mail address: jwhuang@mail.ncku.edu.tw (J.-W. Huang).

https://doi.org/10.1016/j.engappai.2018.07.010
Received 4 February 2018; Received in revised form 23 July 2018; Accepted 29 July 2018

We regard that the same word with different neighboring words would have different meanings. The combination of consecutive words should be seen as a meaningful unit. Therefore, we use bi-grams as a basic unit instead of a single word to be input into the paragraph vector. In addition, using bi-grams enlarges the training corpus and includes more information of the order of words via bi-grams. The proposed Bi-PVDBOW method can represent more semantic meanings of the contents. The experimental results show that the proposed method outperforms original paragraph vectors. Finally, we use the bi-grams-based paragraph vector as an input for classifying the security level of emails by artificial neural network. (2) On the issue of imbalanced security labels, we compared different methods. To avoid overfitting and maintain the distribution of features of original data, we apply K-means clustering to do undersampling. The cluster-based majority undersampling can effectively avoid the critical information loss of majority class. (3) To make our service efficient, we build a distributed system when parsing emails and preprocessing text data on Spark (Zaharia et al., 2010) and Hadoop (White, 2012), which utilized the cloud computing technique on multiple machines.

In the experimental design, one real corporation voluntarily uses the system to scrutinize its vast numbers of emails. The goal of the experiments is to validate the performance of the proposed system and its practical value in the real world. The results show the proposed system achieves high true positive rate and F1-scores in predicting unseen emails' security levels and the average processing time is concise. All validation results are good enough, which implies that the proposed email classification system can effectively and expeditiously operate in the corporation and help to control the sensitive information in real time. Furthermore, the effects of undersampling were examined. The experimental results show that the sampled data can competently represent the original data in the majority class. Additionally, the classification results are worse without undersampling for minority class. Importantly, only less than 10% of emails are reported to managers on real corporate data. The results indicate the proposed system can help companies protecting sensitive information in emails in such an effective and efficient way.

The remainder of the paper is organized as follows. Section 2 outlines the previous work related to the techniques used in this study. Section 3 describes the proposed email classification system in detail. Section 4 shows the experiments aimed at evaluating the prediction results generated by the system. Finally, the conclusions are given in Section 5.

2. Related work

2.1. Security level classification for documents

Security-level labeling of a document is a document classification problem. The purpose of document classification is to assign predefined labels to a new material that is not classified (Joachims, 1998). We have to first transform the textual data into a relational and analytical form. Then the new representation of documents can be fed into a classifier. However, in the security-level classification problem, the amount of data of confidential class is much less than nonconfidential data. We have to deal with the imbalance data problem.

For security-level classifier, Alparslan and Bahsi (2009) used Support Vector Machine (SVM) and Naive Bayes (NB) to classify confidential documents, the SVM on 59 test documents achieved best overall accuracy, 89.83% (53/59). Alparslan et al. (2013) further proposed the SVM-Adaptive Neuro-Fuzzy Inference System that achieved the best overall accuracy, 96.67% (57/59), on 59 test documents. Shakir et al. (2016) proposed an ensemble approach that combines SVM, NB, Decision Trees, and K-Nearest Neighbor. The ensemble approach achieved the classification performance on legal document filtering around 90% Precision, Recall, and F-measure. However, previous works did not deal with the issue of imbalanced data. Therefore, we aim to design a robust classifier learnt from large-scale datasets with short texts and imbalanced classes.

2.2. Document representation

Classification performance of textual data is very relevant to the preprocessing tasks (Han and Kamber, 2006). Effective feature extraction can greatly facilitate machine learning. In this work, the only available information is text contents provided by the corporation. Textual data needs to be formatted in a relational and analytical form. In some works (Alparslan and Bahsi, 2009; Alparslan et al., 2013; Shakir et al., 2016), Term Frequency–Inverse Document Frequency (TF–IDF) is used to represent text-based contents. Using this representation, each of distinct term in the document set is a dimension of the TF–IDF representation. However, TF–IDF leads to a very high-dimensional representation. For example, the dataset used by Alparslan and Bahsi (2009) and Alparslan et al. (2013) has only 222 documents but includes over 2.5 million of words. Therefore, TF–IDF representation may not be appropriate to deal with emails.

This necessitates the formulation of a document representation to replace the article (Jain and Yu, 1998). Numerous methods have been devised for the retrieval of information from text. Word frequency, TF–IDF, is the most common way to retrieve intelligence from text (Salton and Buckley, 1988). Topic models, such as Latent Dirichlet Allocation, LDA (Blei et al., 2003), are statistical models used to discover the topic of a document by extracting latent topical information from the document to generate a document-topic distribution map. LDA is used to derive the topic distribution of a document. In some cases, a distributed vector can be used as a representation of the document.

However, those methods are hard to extract the semantic meaning of words in the document. The neural network is used to learn word vectors for the representation of documents and sentences, in a process referred to as word embedding (Mikolov et al., 0000; Mikolov, 0000; Mikolov et al., 2013). Word vectors were derived from the neural probability language model proposed in Bengio et al. (2003). In the neural probability language model, each input word is represented by a vector and then concatenated or averaged to predict subsequent words in the text. Probability prediction is transformed into a multi-class classification with the architecture of two-layer neural network.

Paragraph vectors (Le and Mikolov, 2014) is an extension of word vectors (Mikolov et al., 0000) to construct embeddings from entire documents using a two-layer neural network, which differ from word vectors by the fact that the input includes a paragraph id. The Skipgram and DBOW of word vector model can be used to compute the paragraph vector by adding a paragraph id. The first architecture can be used to predict the word immediately after the training word. Following the training process, paragraph vectors and word vectors are unique. It operates like a record of missing contents, and is therefore referred to as Distributed Memory model of Paragraph Vector, PVDM (Le and Mikolov, 2014). The second architecture is the Distributed Bag of Words version of Paragraph Vector, PVDBOW (Le and Mikolov, 2014), which does not consider the order of the words. It aims to predict the presence of words in the document. In each training iteration, words from the text window are randomly sampled in order to formulate a classifier capable of predicting the words in a paragraph vector. Paragraph vectors can be used to deal with text of any length.

2.3. Preprocessing methods of imbalanced data

Machine learning-based classifiers learn the characteristics of data by minimizing the error rate. However, the results make sense only if each class of data is balanced. Sampling is one approach to the balancing of data. There are two types of sampling: over-sampling and under-sampling. The over-sampling approach involves increasing the amount of minority class data (Han et al., 2005). The authors oversampling the minority class until the amounts of data in each class balanced. However, over-sampling can lead to problems in the cross validation. Fig. 1 shows two problems of the cross validation due to over-sampling when conducting cross validation. Fig. 1(a) shows over-sampling prior