Design of Multi-View Based Email Classification for IoT Systems via Semi-Supervised Learning

Wenjuan Li[a,b], Weizhi Meng[b,2], Zhiyuan Tan[c], Yang Xiang[d]

[a]*Department of Computer Science, City University of Hong Kong, Hong Kong*
[b]*Department of Applied Mathematics and Computer Science, Technical University of Denmark, Denmark*
[c]*School of Computing, Edinburgh Napier University, United Kingdom*
[d]*Swinburne University of Technology, Australia*
[e]*E-mail address: {weme@dtu.dk}*

Abstract

Suspicious emails are one big threat for Internet of Things (IoT) security, which aim to induce users to click and then redirect them to a phishing webpage. To protect IoT systems, email classification is an essential mechanism to classify spam and legitimate emails. In the literature, most email classification approaches adopt supervised learning algorithms that require a large number of labeled data for classifier training. However, data labeling is very time consuming and expensive, making only a very small set of data available in practice, which would greatly degrade the effectiveness of email classification. To mitigate this problem, in this work, we develop an email classification approach based on multi-view disagreement-based semi-supervised learning. The idea behind is that multi-view method can offer richer information for classification, which is often ignored by literature. The use of semi-supervised learning can help leverage both labeled and unlabeled data. In the evaluation, we investigate the performance of our proposed approach with datasets and in real network environments. Experimental results demonstrate that multi-view can achieve better classification performance than single view, and that our approach can achieve better performance as compared to the existing similar algorithms.

Keywords:
Email Classification, Semi-Supervised Learning, Multi-View Data, Disagreement-based Learning, IoT Security.

1. Introduction

Internet of Things (IoT) represents a network of physical objects containing embedded technologies to sense, communicate and interact with their internal states or the external environment through the Internet connections. With the rapid development of the Internet, sending emails has emerged as an effective and essential way to communicate within various IoT environments for exchanging information. However, due to the rapid increase of IoT devices and nodes, spam or junk emails have become one annoying issue for Internet Service Providers (ISPs) as well as a big threat for IoT security [16, 57, 71]. These suspicious emails can cause various security and privacy issues if they are not timely detected, i.e., spammers could send phishing content as HTML mail, which can carry embedded malicious code or can be enclosed with attachments that contain macro virus. The goal of spam emails is to redirect recipients to pre-built phishing websites that induce users to input their credentials, or automatically infer and collect personal information [42, 47]. As a result, there is a great need for an appropriate security mechanism to classify emails and detect malicious content [24, 65].

In the literature, many supervised machine learning algorithms have been studied to build an email classification system, such as Naive Bayes [28], decision tree [46], k-nearest neighbor [10] and support vector machine (SVM) [1]. Although these supervised methods reported good results in spam identification, they still suffer from several issues in a practical scenario.

- *Demand for diverse labeled data.* Typically, supervised email classification systems require a large number of labeled data (or instances) for classifi-

[1]A preliminary version of this paper appears in Proceedings of the 13th IEEE International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom), pp. 174-181, 2014 [18].

[2]Corresponding author, weme@dtu.dk

er training. In other words, numerous training examples with ground-truth labels should be given in advance. However, only a very small proportion of labeled data is available while most data remains unlabeled in a practical environment.

- *Heavy burden of expert labeling.* Human efforts are extensively demanded for labeling data items to train a supervised learning algorithm. However, due to the high cost of expert labeling, it is very difficult to obtain enough labeled data for classifier training, which significantly hinders the development of supervised email classification systems.

- *Hard to handle unseen data.* In addition, it is very hard to establish an accurate profile for supervised email classification systems, as the number of labeled data is often limited and insignificant. Nowadays, spammers often manipulate an email to bypass a known email system, i.e., the content and structure of spam emails may be quite different from the emails that are used to train a classifier. Therefore, a traditional supervised email classification system cannot detect 'zero-day' emails without appropriate training.

Contributions. Motivated by the challenges above, in this work, we focus on email classification and propose an effective approach by combining both multi-view data and disagreement-based semi-supervised learning. First, we aim to investigate the impact of multi-view data on email classification, which is often ignored by literature. Then, we apply disagreement-based semi-supervised learning for enhancing the performance of spam detection, through leveraging both labeled and unlabeled data. Our contributions of the work can be summarized as follows:

- In this work, we develop an email classification model based on both multi-view data and semi-supervised learning, which adopts two feature sets: *internal feature set (IFS)* and *external feature set (EFS)*. The former contains features that are related to email text (or body), while the latter mainly contains features that are related to routing and forwarding.

- In addition, we revise and deploy a disagreement-based semi-supervised learning algorithm to automatically leverage both labeled and unlabeled data during email classification. This algorithm can make a label decision by means of either "Average of Probabilities" or "Majority Voting". These two methods are also compared in the evaluation.

- To investigate the performance, we first evaluated our proposed classification approach with two datasets: a public dataset and a real (private) dataset, respectively. Then we collaborated with an IT organization and evaluated our approach in a real network environment. Experimental results indicate that our approach can achieve better classification performance as compared to several similar algorithms.

The remaining parts are organized as follows. In Section 2, we review related research regarding the application of machine learning in email classification. Section 3 describes our proposed email classification approach, including how to construct multi-view dataset and how the disagreement-based semi-supervised learning algorithm works. Section 4 presents the experimental settings and analyzes the evaluation results. Finally, we conclude our work in Section 5.

2. Related Work

Email classification is considered to be one promising and commonly adopted method to detect spam emails (e.g., in mobile social networks [43]). Many machine learning algorithms have been studied to distinguish the suspicious emails from the legitimate ones, e.g., supervised learning algorithms and semi-supervised learning algorithms.

Supervised learning algorithms. In the literature, numerous supervised machine learning algorithms have been studied, such as Naive Bayes, decision tree, k-nearest neighbor (KNN), Support Vector Machine (SVM), ensemble learning, etc.

For example, Marsono *et al.* [28] proposed a hardware architecture for a Naive Bayes classifier in the context of email classification for spam control. They particularly presented a word-serial Naive Bayes classifier architecture that utilizes the Logarithmic Number System (LNS) to reduce the computational complexity and for non-iterative binary LNS recoding using a look-up table approach. The experiment showed that their approach could handle large number of emails in second. Meizhen *et al.* [55] proposed a spam-behavioral recognition model and developed a Fuzzy Decision Tree based spam filter system, which computed Information Gain to analyze and select behavior features of emails. Then, Shi *et al.* [46] proposed a novel classification method based on decision tree and introduced an ensemble learning to identify spam emails. The evaluation results on a public dataset indicated that the proposed

method generally outperformed benchmark techniques such as C4.5, Naive Bayes, SVM and KNN.

Regarding KNN and SVM, Firte *et al.* [10] presented an approach for spam detection filters. They particularly developed an offline application that employed the k-Nearest Neighbor (kNN) algorithm and a pre-classified email dataset for the learning process. During the evaluation, this system could perform a constant update to the dataset and the list of most frequently words that appear in the messages. Drucker *et al.* [8] studied the use of support vector machines in classifying emails as spam or legitimate by comparing it to three classification algorithms: Ripper, Rocchio, and boosting decision trees. These four algorithms were tested on two different datasets, where SVM could perform the best when handling binary features. Later, Sculley and Wachman [45] firstly showed that online SVMs indeed gave state-of-the-art classification performance on online spam filtration on large benchmark datasets. They showed that nearly equivalent performance would be achieved by a Relaxed Online SVM (ROSVM) at greatly reduced computational cost. Their results are experimentally verified on email spam, blog spam, and splog detection tasks.

Later, Zhan *et al.* [62] proposed a stochastic learning method to model abnormal emails using weak estimators in a dynamic environment. A multivariate Bernoulli Naive Bayes (NB) classifier was employed in the training phase. The experimental results demonstrate the feasibility and effectiveness of detecting anomalous emails. El-Alfy and Abdel-Aal [9] investigated the application of Group Method of Data Handling based inductive learning approach in detecting spam messages by automatically identifying content features, which can effectively distinguish spam from legitimate emails. Compared with several algorithms like neural networks and Naive Bayes, their approach can provide better spam detection accuracy with false-positive rates as low as 4.3% and also require shorter training time. Ouyang *et al.* [41] conducted a large scale empirical study regarding the effectiveness of using packet and flow features to detect email spam at an organization, based on decision tree and Rulefit. Several other related works can be found in [11, 25, 19, 52, 58, 59, 67, 70, 66].

Semi-Supervised learning algorithms. As supervised learning requires a large number of labeled data, semi-supervised learning has been developed to leverage unlabeled data as well as labeled data for classification.

For instance, Cheng and Li [4] proposed a combined SVM and semi-supervised classifier for increasing the classification accuracy. The SVM is trained with labeled public domain emails aiming to classify a user's emails, while the semi-supervised classifier takes these emails as the training set and propagates the label information to other unlabeled emails by exploiting their distribution in feature space. Then, they [6] further proposed a semi-supervised classifier ensemble aiming to label a users' emails and facilitate the tuning process in an efficient way. This semi-supervised ensemble was validated to help SVM classify users' emails with high accuracy. Gao *et al.* [12] proposed a semi-supervised approach, called regularized discriminant EM algorithm (RDEM), to detect image spam emails. Compared with fully supervised learning algorithms, they found the cost was too high for fully supervised learning to frequently collect sufficient labeled data for training. By contrast, their approach could leverage small amount of labeled data and large amount of unlabeled data for identifying spams and training a classification model simultaneously. Later, Whissell and Clarke [60] considered a specific scenario for semi-supervised spam filtration: that is, when a large amount of training data is available, but only a few true labels can be obtained for that data. They thus presented two spam filtering approaches for such scenario, both starting with a cluster of training emails. In the evaluation, their approach could perform better than those previously published state-of-the-art semi-supervised approach on small-sample spam filtration. Several other related studies about semi-supervised learning in email classification can be referred to [39, 31, 61, 64], and some surveys regarding the spam filtering can be referred to [3, 50, 54].

Semi-supervised learning has proven its capability of detecting spam emails. In the literature, however, we find that very limited research efforts give attention to the use of multi-view data in the field of email classification. In our previous work [18], we aimed to make up this gap and propose an effective classification model by combining both multi-view data and disagreement-based semi-supervised learning. In this work, we extend our previous work and investigate its impact on an IoT environment. In particular, we collaborated with a practical IT organization and tested the approach in a real network environment. We further discuss the open challenges and some limitations in this area.

3. Our Proposed Approach

In this section, we detail the proposed email classification model, including how to construct multi-view data and how the disagreement-based semi-supervised learning algorithm works.

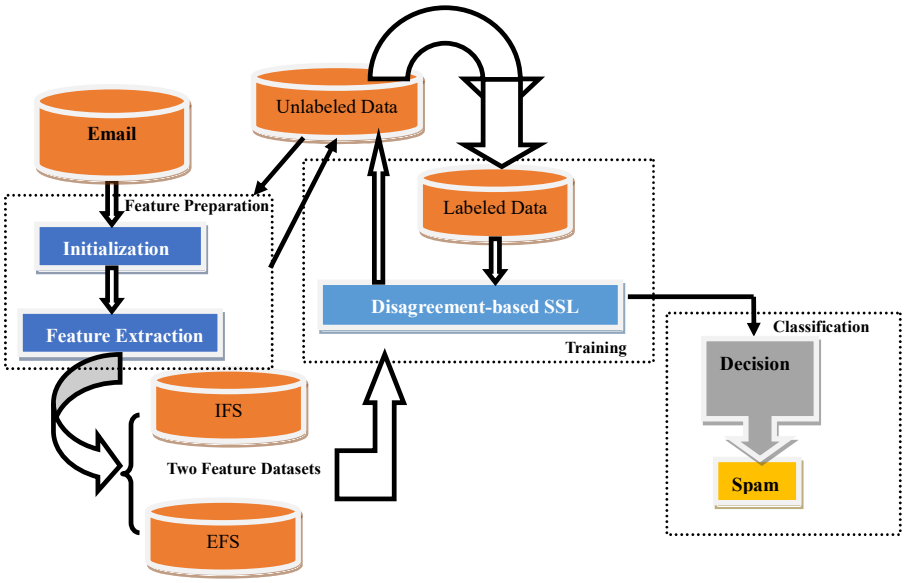

Figure 1: The architecture of our proposed email classification model based on multi-view data and disagreement-based semi-supervised learning.

Table 1: The construction of two feature sets in our email classification model.

Numbers	*Internal Feature Set (IFS)*	*External Feature Set (EFS)*
1	subject length	the number of receipts
2	message size	the number of replies
3	the number of attachments	the level of importance
4	type of attachments	the frequency of sending emails
5	size of attachments	the frequency of receiving emails
6	the number of words in the subject	the name length of senders
7	the number of words in the message	
8	the number of embedded images	

3.1. Email Classification Model

The major purposes of our classification model are twofold: 1) extracting appropriate features from each email that can be handled by classifiers, and constructing two-attribute datasets according to the concept of multi-view; 2) applying a disagreement-based semi-supervised learning algorithm to mark and leverage unlabeled data in an automatic way. The high-level architecture of our proposed email classification model is depicted in Figure 1.

There are three major phases: *feature preparation*, *training* and *classification*. In the first phase, *initialization* is responsible for preprocessing all incoming email messages into our defined common format in order to make incoming emails feasible to be handled by a machine learning classifier (i.e., an email will be represented by a set of features). Then the process of *feature extraction* collects these common features and converts them into two attribute sets: an *internal feature set (IFS)* and an *external feature set (EFS)*. The *IFS* contains attributes that are related to email content (or body), while the *EFS* consists of the ones that are related to routing and forwarding.

In the *training* phase, the implemented disagreement-based semi-supervised learning algorithm can establish classification models by using labeled multi-view instances, and automatically label and leverage unlabeled data. In the last phase, a decision can be made by classifying email messages to either spam or legitimate emails. It is worth noting that all *unlabeled data* (as shown in Figure 1) will be standardized into the common format by passing through the first phase of *feature preparation* in order to facilitate the compatibility of different machine learning classifiers.

3.2. Multi-View Data Construction

Several research studies like [48, 56, 68] in the area of machine learning have shown that multi-view data

can be used to improve the performance of a classifier. In particular, Mao *et al.* [26] applied multi-view to intrusion detection and showed that the use of multi-view data can provide a lower error rate than the use of a single-view data.

In the literature, however, we notice that very few research studies give attention to the use of multi-view data in the field of email classification. To investigate this issue, one of our goals is to explore the impact of multi-view data on email classification. In this part, we introduce how to construct two multi-view feature sets (namely *IFS* and *EFS*) for common emails including those from existing IoT systems. The detailed features are summarized in Table 1.

- *Internal Feature Set (IFS).* These features are relevant to email content or body such as *subject length*, *message size*, *the number of attachments*, *type of attachments*, *size of attachments*, *the number of words in the subject*, *the number of words in the message* and *the number of embedded images*.
- *External Feature Set (EFS).* Different from *IFS*, *EFS* is relevant to email routing and forwarding such as *the number of receipts*, *the number of replies*, *the level of importance*, *the frequency of sending emails*, *the frequency of receiving emails* and *the name length of senders*.

Feature selection and extraction. Some features like *subject length*, *message size*, *size of attachments* and *the number of words in the message* have ever been studied in several research studies like [16, 27] and in public spam datasets like [7]. These studies have proven the feasibility of using these features to describe an email. Based on the above features, in this work, we adopted the above 14 features with two attribute sets to characterize an email. This particular method of data construction makes our work different from most existing work. In real deployment, we identify that the features above can be easily captured and computed by means of current email technique (i.e., route tracking and content recording).

Multi-view. In the literature, we identify that most research explored the issue of email classification using only one attribute data and few studies discuss the multi-view method. Some related work can be referred to [5, 17, 23, 51]. This is because single-view is much straightforward than multi-view data. Motivated by some interesting results like [26], in this work, we attempt to construct a two-view data by using the above proposed features and investigate the influence on email classification.

To better describe our task, let A and B denote two views and (<a,b>, c) denote a labeled example, where $a \in A$ and $b \in B$ are two portions of the example, and c is the label where let 0 denote negative classes and 1 denote positive classes. We assume that there are two functions f_A and f_B over A and B, such that $f_A(a) = f_B(b) = c$. This means that each example is associated with two attributes where each contains enough information for determining the label of the example [68]. Thus, if given k examples, we can have a labeled dataset: $(< a_k, b_k >, c_k)$ (k=1,2,..., c_k *is known*). Let $U = (< a_i, b_i >, c_i)$ ($i = 1, 2...,$ c_i *is unknown*) denote a large number of unlabeled data, our task is to train a classifier to classify new examples.

3.3. Disagreement-based Semi-Supervised Learning

Disagreement-based semi-supervised learning can provide a mechanism to allow classifiers to be trained by different views. The learning process can be treated as a kind of ensemble learning. In addition, semi-supervised learning can refer to either transductive learning or inductive learning. The former attempts to infer the correct labels for the given unlabeled data whereas the latter aims to infer the correct mapping. In practice, a semi-supervised learning algorithm often uses transduction or induction interchangeably.

The goal of disagreement-based semi-supervised learning is to generate multiple learners, enable them collaborating to exploit unlabeled examples, and maintain a large disagreement between the base learners. Regarding the concept of multi-views, we can generate multiple learners using these multi-views and then utilize the multiple learners to start disagreement-based semi-supervised learning. To our knowledge, the co-training algorithm proposed by Blum and Mitchell [2] is the first work by implementing this concept. They assumed that the data has two sufficient and redundant views (i.e., attribute sets), where each view is sufficient for training a strong learner and the views are conditionally independent to each other given the class label.

To better explain the disagreement-based semi-supervised learning, let L and U denote a labeled dataset and an unlabeled dataset respectively, assuming that $L = \{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$ and $U = \{(< x_1^{'}, y_1^{'} >, c_1^{'}), (< x_2^{'}, y_2^{'} >, c_2^{'}), ..., (< x_n^{'}, y_n^{'} >, c_m^{'})\}$. By presenting L and U to a learning algorithm in constructing a function $X \rightarrow Y$, we can predict the labels of unseen data by using this function (where X and Y presents the input space and output space respectively, $x_i, x_j^{'} \in X, i = 1, 2, ..., |n|, j = 1, 2, ..., |m|$). By considering multi-views, L and U can be represented as $L = \{(< x_1, y_1 >, c_1), (< x_2, y_2 >, c_2), ..., (< x_n, y_n >, c_n)\}$

and $U = \{(< x_1^{'}, y_1^{'} >, c_1^{'}), (< x_2^{'}, y_2^{'} >, c_2^{'}), ..., (< x_n^{'}, y_n^{'} >, c_m^{'})\}$ respectively.

As mentioned earlier, it is noticed that several multi-view learning algorithms require independent and redundant views. Unfortunately, such a requirement can hardly be met in most scenarios [69]. In this work, we employ a method of disagreement-based co-training (ensemble) [49], which does not require independent and redundant attributes, but use multiple base classifiers with different learning algorithms on different subsamples of original labeled data.

Specifically, each classifier h is first trained on the original labeled data. Ensembles H are then established by means of all classifiers except one (eh) to search for a subset of high confidence unlabeled data. These ensembles estimate the error rate for each classifier from the agreement among the classifiers. Later, a subset of U is selected by eh for h. Data that can improve the error over a pre-defined threshold are added to the labeled training dataset. In this case, each classifier has its own training dataset. Note that data that is labeled for the classifier is not deleted from the unlabeled dataset. The above training process will be repeated until there are no more data can be labeled to improve the performance of any classifier. An outline of this co-training is shown as below:

- Initialization: given L, U, H;
- For each iteration i:
  - Finding error rate for component classifier based on disagreement among classifiers;
  - Assigning labels to unlabeled instances based on agreement among ensembles;
  - Sampling high-confidence examples for component classifier;
  - Building component classifier based on newly-labeled and original labeled instances;
  - Iteration end.
  - Controlling the error rate for each component classifier and update the ensemble.
- Generating final hypothesis.

The specific co-training algorithm can be referred to [49], but differently, we employ OLTV method [68] to generate L and U for the co-training which can help generate a more reliable dataset. The OLTV method assumes that if two sufficient views are conditionally independent given the class label, the most strongly correlated pair of projections should be in accordance with the ground truth. The specific algorithm of OLTV is described in Table 2. To label an unlabeled data, we employ two voting approaches: "Average of Probabilities" [49] and "Majority Voting" [15].

Table 2: The OLTV algorithm.

Process:
1. $L_P \leftarrow seed$, $L_N \leftarrow \emptyset$
2. Identify all pairs of correlated projections, obtaining α_i, β_i *and* λ_i.
3. For $j = 0, 1, 2, ..., l-1$ do $Project < x_i, y_i >$ into the m pairs of correlated projections.
4. For $j = 1, 2, ..., l-1$ do compute ρ_i
5. $P \leftarrow argmax_{\gamma+}(\rho_i)$, $N \leftarrow argmin_{\gamma-}(\rho_i)$
6. For all $< x_j, y_j > \in P$ do $L_P \leftarrow L_P \cup (< x_j, y_j >, 1)$
7. For all $< x_j, y_j > \in N$ do $L_N \leftarrow L_N \cup (< x_j, y_j >, 0)$
8. $L \leftarrow L_P \cup L_N$, $U \leftarrow U - (P \cup N)$

Output: L, U.

For the method of 'Average of Probabilities", we suppose $Y = (y_1, y_2, ..., y_m)$ to be the class labels and there are totally N classifiers. This voting method for predicting the new example x can be computed as:

$$arg\ max(\frac{1}{N}\sum_{i=1}^{N} p_i(y_m|x)) \quad (1)$$

For the method of "Majority Voting", the maximum number of classifiers is considered as a major rule, which means that the majority of the classifiers should be agreed to assign a label to one unlabeled data.

4. Evaluation

In this section, we evaluate our proposed email classification model using two datasets (a public dataset and a real dataset) and in a real network environment. The use of two datasets attempts to investigate the performance of disagreement-based learning algorithm and the impact of multi-view data. The evaluation in a real network environment aims to explore the real performance of our approach. Below are the metrics adopted in the evaluation.

- *Area under an ROC curve (AUC).* This is an important metric used for comparing various classifiers. It represents the expected performance as a single scalar in which the larger the AUC, the better the experiment is as predicted the existence of the classification [44].

- *False positive rate (FPR).* This metric indicates the rate of classifying a legitimate email as a spam.
- *False negative rate (FNR).* This metric indicates the rate of classifying a spam email as a legitimate one.
- *Classification accuracy.* This metric indicates the rate of correctly classifying both spam and legitimate emails.

4.1. Experiment1

In this experiment, we aim to explore the performance of disagreement-based semi-supervised learning algorithm as compared to several traditional supervised learning classifiers such as Naive Bayes, IBK (with $k = 3$), J48 and SMO. With the purpose of avoiding any implementation bias, all classifiers are extracted from the WEKA platform [53].

To explore the performance, we adopted a publicly available spam email dataset, called *SPAM E-mail Dataset* [7], which contains 58 attributes and a total of 4601 emails (1813 spam emails and 2788 legitimate emails). To evaluate the disagreement-based semi-supervised learning algorithm, we divided this dataset into two parts: labeled data and unlabeled data, where the unlabeled data consists of 600 instances that are randomly selected from the original dataset. Then we compared three classifiers: Naive Bayes, IBK and J48 in the disagreement-based learning and set the value of pre-defined threshold to 0.75 for all classifiers. The disagreement-based semi-supervised learning algorithm was tested by 60 iterations based on "Majority Voting". The experimental results are shown in Table 3.

Table 3: Classification comparison in *Experiment1*.

Algorithm	FPR	FNR	Classification Accuracy
Naive Bayes	0.169	0.248	0.765
SMO	0.142	0.223	0.783
IBK	0.134	0.215	0.792
J48	0.113	0.187	0.823
Our algorithm	0.092	0.101	0.884

It is found that the disagreement-based semi-supervised learning algorithm could outperform other supervised learning algorithms in the aspects of false positives, false negatives and classification accuracy. For example, J48 achieved the best classification accuracy of 0.823 among the supervised learning classifiers while the disagreement-based semi-supervised learning could increase the classification accuracy to 0.884. Our approach can also achieve lower false rates (for both FPR and FNR) than the others, i.e., our approach achieved 0.092 and 0.101 for FPR and FNR. These results indicate that semi-supervised learning can overall enhance the classification capability of spam detection by leveraging both labeled and unlabeled data.

Table 4: Comparison of classification accuracy using "Average of Probabilities" and "Majority Voting".

Voting Methods	60 Iterations	100 Iterations
Average of Probabilities	0.852	0.904
Majority Voting	0.857	0.913

4.2. Experiment2

As there is no publicly adopted dataset for multi-view data in security, in this experiment, we evaluated the performance of our email classification model by constructing a private dataset based on our defined features (see Table 1). The dataset is mainly comprised of 7133 emails recorded from two recognized institutes, which was similarly divided into two parts: labeled dataset and unlabeled dataset via a random selection process. The unlabeled dataset contains 2300 instances selected from the private dataset, while the remaining data was manually labeled by three security officers from the institutes. We also used the same classifiers: Naive Bayes, IBK and J48 in the disagreement-based SSL and set the value of pre-defined threshold to 0.75 for all classifiers.

"Average of Probabilities" versus "Majority Voting". To explore the performance between these two voting methods, we compared the classification accuracy after 60 and 100 iterations, respectively. Table 4 shows that these two voting methods can achieve very similar classification accuracy, but the method of "Majority Voting" could still reach a better accuracy rate for our classification model.

Multi-view versus single-view. To investigate the impact of multi-view data on email classification, we compared our approach with the single-view EM semi-supervised learning [40]. As shown in Table 1, it is worth noting that all features have to be used to train the EM semi-supervised learning algorithm as a single view dataset. Our approach adopted "Majority Voting" in this experiment, and detailed results are shown in Figure 2.

It is found that our approach with multi-view data could outperform the use of single-view data by gradually increasing the classification accuracy. Also, it is observed that our approach could improve the classification accuracy significantly after a few training iterations. For instance, after 60 iterations, our approach

Table 5: Comparison of classification results in *Experiment2*.

Algorithm	Classification Accuracy	AUC
Naive Bayes	0.702	0.761
SMO	0.748	0.779
IBK	0.773	0.796
J48	0.785	0.823
Our algorithm (60 iterations)	0.857	0.913
Standard Co-Training (60 iterations) [2]	0.822	0.897
Co-EM (60 iterations) [40]	0.831	0.902

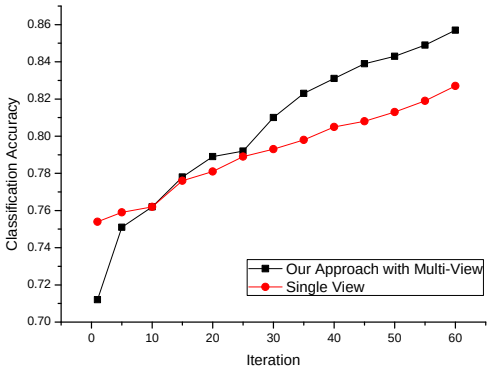

Figure 2: The comparison results of classification accuracy regarding multi-view and single view.

could increase the classification accuracy by nearly 3% as compared to the EM semi-supervised learning with single-view data.

Table 5 further shows a comparison of classification accuracy and AUC between our approach and several supervised learning algorithms. Note that all features would be used to train these supervised learning algorithms as a single-view dataset. It is found that our classification model could reach a better result than other classifiers, through combining multi-view data and semi-supervised SSL. In addition, these supervised learning algorithms only achieved a rate less than 0.8 for both classification accuracy and AUC, reflecting the difficulty of identifying spam emails in real scenarios.

Multi-view algorithm comparison. To further investigate the performance among multi-view disagreement-based SSL algorithms, we conducted a comparison by involving two more algorithms: *Standard Co-Training* [2] and *Co-EM* [40]. Table 5 shows that our algorithm could still outperform the other two algorithms in the aspects of classification accuracy and AUC, i.e., our algorithm achieved an accuracy of 0.857, which is higher than the other two with 0.822 and 0.831.

Overall, these experimental results demonstrate that multi-view data can help achieve better classification accuracy and AUC as compared to the use of single-view data, and that our proposed email classification approach is effective in email classification.

4.3. Experiment3

In this experiment, we collaborated with an IT organization to explore our approach in a collaborative network environment, which includes 42 IoT nodes with laptop, PC, and different sensors. Figure 3 shows the high-level network architecture, in which several nodes can communicate with others as a subgroup. A firewall is deployed between the internal nodes and the Internet. We randomly selected five places to deploy our email classification approach and monitor the performance, including four IoT nodes plus the firewall. The experiment was run for a week under the support from the security managers from the participating organization.

Table 6 shows the average classification accuracy and AUC. The major observations are discussed as follows:

- As compared with the traditional supervised learning, our approach could reach a much better result in the aspects of classification accuracy and AUC. For example, all supervised classifiers could only reach classification accuracy below 0.84, but our approach could achieve a rate above 0.93.

- For the other two multi-view disagreement-based SSL algorithms algorithms, they could perform better than all the supervised classifiers, but our approach could still reach higher accuracy and AUC, i.e., our approach could provide 0.932 and 0.953 regarding classification accuracy and AUC; while standard co-training and Co-EM could only achieve 0.882 & 0.887, and 0.897 & 0.912, respectively.

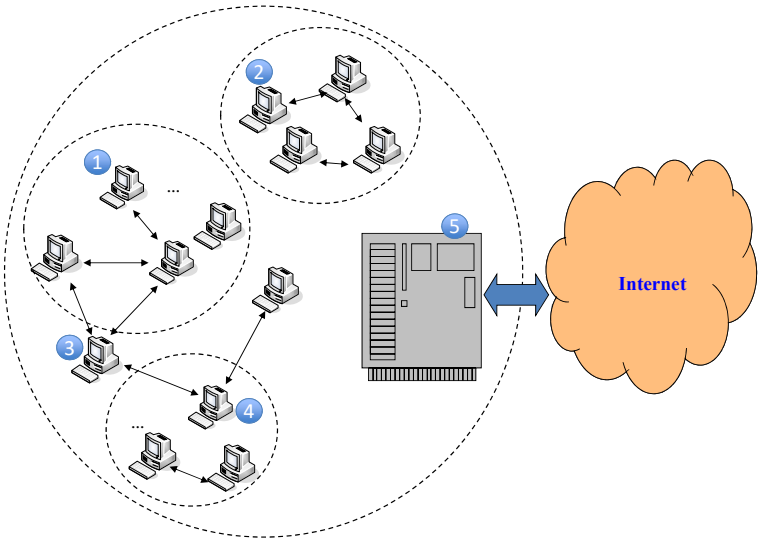

Figure 3: The high-level network architecture in a real organization.

Table 6: Results of average classification accuracy and AUC in *Experiment3*.

Algorithm	Classification Accuracy	AUC
Naive Bayes	0.723	0.742
SMO	0.768	0.789
IBK	0.812	0.823
J48	0.834	0.855
Our algorithm	0.932	0.953
Standard Co-Training [2]	0.882	0.887
Co-EM [40]	0.897	0.912

4.4. Further Discussion

In the evaluation, we present encouraging results achieved by our proposed email classification model. There are several open challenges may affect the performance of email classification.

Multi-view data construction. Based on the structure of an incoming email, some more features can be considered for *IFD* and *EFD*, such as sequence of words and sequence of incoming emails (Temporal Features [14]). It is an interesting topic to explore the stability of classification accuracy under different construction ways of two-feature datasets.

Unlabeled data selection. In the evaluation, we constructed a dataset with unlabeled instances by means of a random selection, which means that it is possible to produce certain "weak" unlabeled dataset or "weak" labeled dataset. This situation may also affect the classification accuracy. We plan to investigate this issue in our future work.

Other learning algorithms. With the recent development of machine learning, more learning algorithms can be considered in our future work, e.g., deep learning [22]. It is also an interesting direction to study some well-built classifier in malware detection [20] and biometric authentication [21, 30, 33, 35, 36].

5. Conclusion

Suspicious emails are a big threat for IoT security. To mitigate this issue, email classification is one basic and important solution. In the literature, many supervised learning classifiers have been studies; however, several challenges remain unsolved in practice such as *requirement of large labeled data*, *heavy burden of expert labeling* and *hard to handle unseen data*.

In this work, we develop an effective email classification model for IoT systems, by combining both multi-view data and disagreement-based semi-supervised learning. For the multi-view data, we construct a two-

view dataset: *internal feature set (IFS)* and *external feature set (EFS)*. The former contains features that are related to email text (or body) while the latter mainly contains features that are related to routing and forwarding. The objective of disagreement-based semi-supervised learning is to leverage both labeled and unlabeled data. In the evaluation, we conducted three major experiments to investigate the performance of our approach, with two datasets and in a real network environment. The experimental results demonstrate that our multi-view construction can improve the accuracy of classifying emails as compared to the use of single-view data, and that our algorithm is effective in practice as compared to several existing multi-view semi-supervised learning algorithms.

To the best of our knowledge, our work is an early effort in discussing the use of multi-view data in email classification. There are many possible topics for our future work, which could include exploring the performance of using other semi-supervised learning algorithms in our proposed model and providing a comparison study. Future work could also include investigating how to systematically construct an appropriate multi-view dataset for the emails from different IoT systems and explore whether there is an optimal construction. It is also an interesting topic to explore the effectiveness of other filtration mechanisms in this area [13, 29, 32, 34, 38].

Reference

[1] O. Amayri and N. Bouguila, "A study of spam filtering using support vector machines," *Artificial Intelligence Review* 34(1), 73-108, 2010.

[2] A. Blum and T. Mitchell, "Combining labeled and unlabeled data with co-training," *Proceedings of the 11th Annual Conference on Computational Learning Theory*, pp. 92-100, 1998.

[3] G. Caruana and M. Li, "A survey of emerging approaches to spam filtering," *ACM Computing Surveys* 44(2), pp. 1-27, 2008.

[4] V. Cheng and C.-H. Li, "Combining supervised and semi-supervised classifier for personalized spam filtering," *Proceedings of PAKDD*, LNAI 4426, pp. 449-456, 2007.

[5] X. Chen, H. Yin, F. Jiang, and L. Wang, "Multi-view dimensionality reduction based on Universum learning," *Neurocomputing* 275, pp. 2279-2286, 2018.

[6] V. Cheng and C.-H. Li, "Personalized spam filtering with semi-supervised classifier ensemble," *Proceedings of the 2006 International Conference on Web Intelligence (WI)*, pp. 195-201, 2006.

[7] SPAM E-mail Dataset (Accessed on 8 September, 2013): `http://web.cs.wpi.edu/~cs4445/b12/Datasets/spambase.arff.`

[8] H. Drucker, D. Wu, and V.N. Vapnik, "Support vector machines for spam categorization," *IEEE Transactions on Neural Networks* 10(5), 1048-1054, 1999.

[9] E.M. El-Alfy and R.E. Abdel-Aal, "Using GMDH-based networks for improved spam detection and email feature analysis," *Applied Soft Computing* 11, pp. 477-488, 2011.

[10] L. Firte, C. Lemnaru, and R. Potolea, "Spam detection filter using KNN algorithm and resampling," *Proceedings of the 6th International Conference on Intelligent Computer Communication and Processing (ICCP)*, pp. 27-33, 2010.

[11] D.M. Freeman, "Using Naive Bayes to detect spammy names in social networks," *Proceedings of the ACM Conference on Computer and Communications Security (CCS)*, pp. 3-12, 2013.

[12] Y. Gao, M. Yang, and A. Choudhary, "Semi supervised image spam hunter: A regularized discriminant EM approach," *Proceedings of ADMA 2009*, LNAI 5678, pp. 152-164, 2009.

[13] L. Jiang, Y. Cheng, L. Yang, J. Li, H. Yan, and X. Wang, "A Trust-Based Collaborative Filtering Algorithm for E-Commerce Recommendation System," *Journal of Ambient Intelligence and Humanized Computing*, pp. 1-12, 2018.

[14] S. Kiritchenko, S. Matwin, and S. Abu-hakima, "Email classification with temporal features," *Proceedings of the International Intelligent Information Systems (IIS)*, pp. 523-533, 2004.

[15] J. Kittler, M. Hatef, R.P. Duin, and J. Matas, "On Combining Classifiers," *IEEE Transactions on Pattern Analysis and Machine Intelligence* 20(3), 226-239, 1998.

[16] R. Islam and Y. Xiang, "Email Classification Using Data Reduction Method," *Proceedings of the 5th International ICST Conference on Communications and Networking in China*, pp. 1-5, 2010.

[17] K.-Y. Lee and J.-Young Sim, "Stitching for Multi-View Videos With Large Parallax Based on Adaptive Pixel Warping," *IEEE Access* 6, pp. 26904-26917, 2018.

[18] W. Li, W. Meng, Z. Tan, and Y. Xiang, "Towards Designing An Email Classification System Using Multi-View Based Semi-Supervised Learning," *Proceedings of the 13th IEEE International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom)*, pp. 174-181, 2014.

[19] W. Li, W. Meng, L.-F. Kwok, and H.H.S. Ip, "Enhancing Collaborative Intrusion Detection Networks Against Insider Attacks Using Supervised Intrusion Sensitivity-Based Trust Management Model," *Journal of Network and Computer Applications*, vol. 77, pp. 135-145, 2017.

[20] Y. Li, G. Wang, L. Nie, Q. Wang, W. Tan, "Distance metric optimization driven convolutional neural network for age invariant face recognition," *Pattern Recognition* 75, pp. 51-62, 2018.

[21] J. Li, L. Sun, Q. Yan, Z. Li, W. Srisa-an, and H. Ye: "Significant Permission Identification for Machine-Learning-Based Android Malware Detection," *IEEE Trans. Industrial Informatics* 14(7): 3216-3225 (2018)

[22] Y. Liu, J. Ling, Z. Liu, J. Shen, and C. Gao, "Finger vein secure biometric template generation based on deep learning," *Soft Comput.* 22(7), pp. 2257-2265, 2018.

[23] Y. Liu, C. Jiang, and H. Zhao: Using contextual features and multi-view ensemble learning in product defect identification from online discussion forums. *Decision Support Systems* 105, pp. 1-12, 2018.

[24] L. Liu, O. De Vel, Q.-L. Han, J. Zhang, and Y. Xiang, Detecting and Preventing Cyber Insider Threats: A Survey, IEEE Communications Surveys and Tutorials, vol. 20, no. 2, pp. 1397-1417, 2018.

[25] C. Lopes, P. Cortez, P. Sousa, M. Rocha, and M. Rio, "Symbiotic filtering for spam email detection," *Expert Systems with Applications* 38, pp. 9365-9372, 2011.

[26] C.-H. Mao, H.-M. Lee, D. Parikh, T. Chen, and S.-Y. Huang, "Semi-supervised co-training and active learning based approach for multi-view intrusion detection," *Proceedings of the 2009 ACM symposium on Applied Computing (SAC)*, pp. 2042-2048, 2009.

[27] S. Martin, A. Sewani, B. Nelson, K. Chen, and A.D. Joseph, "Analyzing Behaviorial Features for Email Classification," *Proceedings of the 2005 Annual Collaboration, Electronic messaging, Anti-Abuse and Spam Conference (CEAS)*, pp. 1-8, 2005.

[28] M.N. Marsono, M.W. El-Kharashi, and F. Gebali, "Binary LNS-based naive Bayes hardware classifier for spam control," *Proceedings of IEEE International Symposium on Circuits and Systems*, pp. 3674-3677, 2006.

[29] W. Meng, "Intrusion Detection in the Era of IoT: Building Trust via Traffic Filtering and Sampling," IEEE Computer, vol. 51, no. 7, pp. 36-43, 2018.

[30] Y. Meng, D.S. Wong, R. Schlegel, and L.-F. Kwok, "Touch Gestures Based Biometric Authentication Scheme for Touchscreen Mobile Phones," *Proceedings of the 8th China International Conference on Information Security and Cryptology (INSCRYPT)*, pp. 331-350, 2012.

[31] Y. Meng, W. Li, and L.F. Kwok. "Enhancing Email Classification Using Data Reduction and Disagreement-based Semi-Supervised Learning," *Proceedings of The 2014 IEEE International Conference on Communications (ICC)*, 2014.

[32] W. Meng and L.-F. Kwok, "Enhancing the Performance of Signature-based Network Intrusion Detection Systems: An Engineering Approach," *HKIE Transactions*, vol. 21, no. 4, pp. 209-222, Taylor & Francis, 2014.

[33] Y. Meng, D.S. Wong, and L.F. Kwok, "Design of Touch Dynamics based User Authentication with an Adaptive Mechanism on Mobile Phones," *Proceedings of the 29th Annual ACM Symposium on Applied Computing: the 13th edition of the Computer Security track (ACM SAC SEC)*, pp. 1680-1687, 2014.

[34] W. Meng, W. Li, and L.-F. Kwok, "EFM: Enhancing the Performance of Signature-based Network Intrusion Detection Systems Using Enhanced Filter Mechanism," *Computers & Security*, vol. 43, pp. 189-204, 2014.

[35] W. Meng, D.S. Wong, S. Furnell, and J. Zhou, "Surveying the Development of Biometric User Authentication on Mobile Phones," *IEEE Communications Surveys & Tutorials*, vol. 17, no. 3, pp. 1268-1293, 2015.

[36] W. Meng, W. Li, and D.S. Wong, "Enhancing Touch Behavioral Authentication via Cost-based Intelligent Mechanism on Smartphones," *Multimedia Tools and Applications*, pp. 1-12, 2018.

[37] W. Meng, Y. Wang, D.S. Wong, S. Wen, and Y. Xiang, "TouchWB: Touch Behavioral User Authentication Based on Web Browsing on Smartphones," *Journal of Network and Computer Applications*, vol. 117, pp. 1-9,2018.

[38] W. Meng, W. Li, C. Su, J. Zhou, and R. Lu, "Enhancing Trust Management for Wireless Intrusion Detection via Traffic Sampling in the Era of Big Data," *IEEE Access*, vol. 6, no. 1, pp. 7234-7243, IEEE, 2018.

[39] M. Mojdeh and G.V. Cormack, "Semi-supervised spam filtering using aggressive consistency learning," *Proceedings of the 33rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pp. 751-752, 2010.

[40] K. Nigam, A. McCallum, and T. Mitchell, "Semi-supervised Text Classification Using EM," In: Chapelle, O., Zien, A., and Scholkopf, B. (eds.) Semi-Supervised Learning. MIT Press: Boston (2005)

[41] T. Ouyang, S. Ray, M. Allman, and M. Rabinovich, "A large-scale empirical analysis of email spam detection through network characteristics in a stand-alone enterprise," *Computer Networks* 59, pp. 101-121, 2014.

[42] S. Peng, A. Yang, L. Cao, S. Yu, D. Xie, "Social influence modeling using information theory in mobile social networks," *Inf. Sci. 379*, pp. 146-159, 2017.

[43] T. Peng, Q. Liu, D. Meng, G. Wang, "Collaborative trajectory privacy preserving scheme in location-based services," *Inf. Sci. 387*, pp. 165-179, 2017.

[44] S. Rosset, "Model selection via the AUC," *Proceedings of the 21th International Conference on Machine Learning (ICML)*, pp. 89-97, 1989.

[45] D. Sculley and G.M. Wachman, "Relaxed Online SVMs for Spam Filtering," *Proceedings of the 30th annual international ACM SIGIR conference on Research and Development in Information Retrieval (SIGIR)*, pp. 415-422, 2007.

[46] L. Shi, Q. Wang, X. Ma, M. Weng, and H. Qiao, "Spam email classification using decision tree ensemble," *Journal of Computational Information Systems* 8(3), pp. 949-956, 2012.

[47] D. Shinder, "E-mail spam: Is it a Security Issue?" (Accessed on 30 September, 2013) `http://www.windowsecurity.com/articles-tutorials/content_security/Email_Spam.html`.

[48] S. Sun, "A survey of multi-view machine learning," *Neural Comput & Applic* 23, pp. 2031-2038, 2013.

[49] J. Tanha, M. van Someren, and H. Afsarmanesh, "Disagreement-Based Co-training," *Proceedings of the 23rd IEEE International Conference on Tools with Artificial Intelligence (ICTAI)*, pp. 803-810, 2011.

[50] G. Tang, J. Pei, and W.-S. Luk, "Email mining: tasks, common techniques, and tools," *Knowledge and Information Systems*, In Press, 2013.

[51] J. Tang, D. Li, Y. Tian, and D. Liu, "Multi-view learning based on nonparallel support vector machine," *Knowl.-Based Syst.* 158, pp. 94-108, 2018.

[52] S.K. Trivedi and S. Dey, "Document Effect of feature selection methods on machine learning classifiers for detecting email spams," *Proceedings of the 2013 Research in Adaptive and Convergent Systems (RACS)*, pp. 35-40, 2013.

[53] The University of Waikato. WEKA-Waikato Environment for Knowledge Analysis. `http://www.cs.waikato.ac.nz/ml/weka/`

[54] D. Wang, D. Irani, and C. Pu, "A study on evolution of email spam over fifteen years," *Proceedings of the 9th IEEE International Conference on Collaborative Computing: Networking, Applications and Worksharing (COLLABORATECOM)*, pp. 1-10, 2013.

[55] M. Wang, Z. Li, and S. Zhong, "A method for spam behavior recognition based on fuzzy decision tree," *Proceedings of the 9th International Conference on Computer and Information Technology (CIT)*, pp. 236-241, 2009.

[56] W. Wang and Z.-H. Zhou, "On multi-view active learning and the combination with semi-supervised learning," *Proceedings of the 25th International Conference on Machine Learning (ICML)*, pp. 1152-1159, 2008.

[57] C. Wang, J. Shen, Q. Liu, Y. Ren, and T. Li: "A Novel Security Scheme Based on Instant Encrypted Transmission for Internet of Things," *Security and Communication Networks*, pp. 1-7, 2018.

[58] S. Wen, Y. Xiang, and W. Zhou: Modeling and Analysis for Thwarting Worm Propagation in Email Networks. Proceeedings of NSS, pp. 763-769, 2013.

[59] S. Wen, W. Zhou, J. Zhang, Y. Xiang, W. Zhou, W. Jia, and C.C. Zou: Modeling and Analysis on the Propagation Dynamics of Modern Email Malware. IEEE Trans. Dependable Sec. Comput. 11(4), pp. 361-374, 2014.

[60] J.S. Whissell and C.L.A. Clarke, "Clustering for semi-supervised spam filtering," *Proceedings of the 8th Annual Collaboration, Electronic messaging, Anti-Abuse and Spam Con-*

ference (CAES), pp. 125-134, 2011.

[61] Y.-S. Wu, S. Bagchi, N. Singh, and R. Wita, "Spam detection in voice-over-IP calls through semi-supervised clustering," *Proceedings of the International Conference on Dependable Systems and Networks (DSN)*, pp. 307-316, 2009.

[62] J. Zhan, B.J. Oommen, and J. Crisostomo, "Anomaly Detection in Dynamic Systems Using Weak Estimators," *ACM Transactions on Internet Technology* 11(1), pp. 1-16, 2011.

[63] M.-L. Zhang and Z.-H. Zhou, "Multi-label learning by instance differentiation," *Proceedings of the 22nd Conference on Artificial Intelligence (AAAI)*, pp. 669-674, 2007.

[64] W. Zhang, D. Zhu, Y. Zhang, G. Zhou, and B. Xu, "Harmonic functions based semi-supervised learning for web spam detection," *Proceedings of the 2011 ACM Symposium on Applied Computing (SAC)*, pp. 74-75, 2011.

[65] J. Zhang, X. Chen, Y. Xiang, W. Zhou, and J. Wu, Robust Network Traffic Classification, IEEE/ACM Transactions on Networking, vol. 23, no. 4, pp. 1257-1270, Aug, 2015.

[66] B. Zhou, Y. Yao, and J. Luo, "Cost-sensitive three-way email spam filtering," *Journal of Intelligent Information Systems* 42(1), pp. 19-45, 2014.

[67] W. Zhou, S. Wen, Y. Wang, Y. Xiang, and W. Zhou: An Analytical Model on the Propagation of Modern Email Worms. Proceedings of TrustCom 2012, pp. 533-540, 2012.

[68] Z.-H. Zhou, D.-C. Zhan, Q. Yang, "Semi-supervised learning with very few labeled training examples," *Proceedings of the 22nd National Conference on Artificial Intelligence (AAAI)*, pp. 675-680, 2007.

[69] Z.-H. Zhou and M. Li, "Tri-training: Exploiting unlabeled data using three classifiers," *IEEE Trans Knowledge and Data Engineering* 17(11), pp. 1529-1541, 2005.

[70] Y. Zhang, S. Wang, P. Phillips, and G. Ji, "Binary PSO with mutation operator for feature selection using decision tree applied to spam detection," *Knowledge-Based Systems* 64, pp. 22-31, 2014.

[71] Y. Zhang, D. Zheng, R.H. Deng, "Security and Privacy in Smart Health: Efficient Policy-Hiding Attribute-Based Access Control," *IEEE Internet of Things Journal* 5(3), pp. 2130-2145, 2018.