A Multi-Resolution-Concentration Based Feature Construction Approach for Spam Filtering

Guyue Mi, Pengtao Zhang and Ying Tan

Abstract— **This paper proposes a multi-resolution-concentration (MRC) based feature construction approach for spam filtering by progressively partitioning an email into local areas on smaller and smaller resolutions. The MRC approach depicts a dynamic process of gradual refinement in locating the pathogens by calculating concentrations of detectors on local areas, and is considered to be able to extract the position-correlated and process-correlated information from emails. Furthermore, A weighted MRC (WMRC) approach is presented by considering the different activity levels of detectors in calculation of concentrations. A generic structure of the MRC model, which mainly contains detector sets construction and multi-resolution concentrations calculation, is designed. The implementations of MRC and WMRC approaches are described in detail. Experiments are conducted on five benchmark corpora using cross-validation to evaluate the proposed MRC model. Comprehensive experimental results suggest that the MRC and WMRC approaches perform far better than the prevalent bag-of-words approach in both performance and efficiency. Compared with the concentration based feature construction approach and local-concentration based feature extraction approach, MRC and WMRC achieve higher accuracy and F_1 measure, which demonstrates the effectiveness of the MRC model. In addition, it is shown that both the MRC and WMRC approaches cooperate well with variety of classification methods, which endows the MRC model with flexible capability in the real world.**

I. Introduction

CHEAP, efficient and easy to use, email has become an indispensable tool in daily communication. However, affect on normal email communication from spam, which is always aimed at commercial promotion or marketing, is getting more and more serious.

Spam is generally defined as unsolicited bulk email (UBE) or unsolicited commercial email (UCE) [1]. According to Symantec report [2], though Rustock, one of the largest botnets in the world, was closed in 2011, spam made up 69% of the total email traffic in January, 2012. The statistics from Commtouch internet threats report [3] demonstrate that the quantity of spam in the first quarter of 2012 decreased a little bit compared with last year, but still 94 billion spam in average were sent every day. Ferris research [4] revealed large amount of spam not only occupied network bandwidth and server storage, but also made users spend time to read and delete them, which resulted in loss of productivity. Furthermore, spam has been used as carrier of malwares, threatening the safety of internet and personal privacy.

Many solutions have been proposed to address the spam problem, and automatically filtering spam is considered to be the most efficient and effective. The main step of spam filtering is email classification, categorizing an email as spam or ham (legitimate email), which contains two phases, namely feature construction and classifier design. Machine learning methods are widely applied for classifier design in spam filtering, such as naive bayes (NB) [5], [6], support vector machine (SVM) [7], artificial neural network (ANN) [8], [9], k-nearest neighbor (k-NN) [10], [11], boosting [12] and so on. Feature construction approaches are utilized for transforming emails into feature vectors, which can be recognized and categorized by classifiers. Since performance of the machine learning method seriously depends on feature vectors constructed by the feature construction approach, research on feature construction approaches has been focused in recent years.

In this paper, a multi-resolution-concentration (MRC) based feature construction approach for spam filtering is proposed, which progressively partitions an email into local areas on smaller and smaller resolutions, and the concentration features are constructed on each local area. The MRC approach depicts a dynamic process of gradual refinement in locating the pathogens by calculating concentrations of detectors on local areas and is considered to be able to extract the position-correlated and process-correlated information from emails. Furthermore, by introducing the different activity levels of detectors, a weighted MRC (WMRC) approach is presented. A generic structure of the MRC model is designed and the detailed implementations of MRC and WMRC are described. Experiments are conducted on five benchmark corpora PU1, PU2, PU3, PUA and Enron-Spam for investigating performance of the MRC and WMRC approaches. Accuracy and F_1 measure are selected as the main criteria in analyzing and discussing the results.

The rest of the paper is organized as follows: Section II introduces the prevalent feature construction approaches. The proposed MRC and WMRC approaches are presented in Section III. Section IV gives the detailed experimental setup and results. Finally, we conclude the paper in Section V.

II. Prevalent Feature Construction Approaches

A. Bag-of-Words

Bag-of-Words (BoW), also known as space vector model, is one of the most widely used feature construction approach-

Guyue Mi, Pengtao Zhang and Ying Tan are with the Key Laboratory of Machine Perception (Ministry of Education) and Department of Machine Intelligence, School of Electronics Engineering and Computer Science, Peking University, Beijing, 100871, China (email: gymi@pku.edu.cn, pengtaozhang@gmail.com, ytan@pku.edu.cn).

This work is supported by the National Natural Science Foundation of China under grants No. 61170057 and 60875080.

Prof. Ying Tan is the corresponding author.

es in spam filtering and can construct discriminative feature vectors [13]. It transforms an email m to a n-dimensional feature vector $\vec{x} = [x_1, x_2, ..., x_n]$ by utilizing a preselected term set $T = [t_1, t_2, ..., t_n]$, where the value x_i is given as a function of the occurrence of t_i in m, depending on the representation of the features adopted. In a binary representation, x_i is equal to 1 when t_i occurs in m, and 0 otherwise. In a frequency representation, x_i is assigned as the number of occurrences of t_i in m. Experimental results in [14] show binary representation and frequency representation have similar performance. Experiments in [7] reveal SVM can achieve best performance when binary representation is adopted.

B. Sparse Binary Polynomial Hashing

Yerazunis utilized Sparse Binary Polynomial Hashing (SBPH) to extract a large amount of different features from email [15]. This method applies an n-term-length sliding window shifting over an email with a step of one term. At each movement, the newest term in the window is retained and the others are removed or retained so that the whole window is mapped to different features. Hence, 2^{n-1} features are extracted at each movement. SBPH performed quite promisingly in terms of classification accuracy in experiments as it could extract enough discriminative features. However, so many features lead to a heavy computational burden and limit its usability.

C. Orthogonal Sparse Bigrams

Siefkes *et al.* proposed Orthogonal Sparse Bigrams (OSB) based on analyzing the relevance of features to reduce the redundancy and complexity of SBPH [16]. The same as SBPH, OSB also utilizes an n-term-length sliding window shifting over an email with a step of one term. However, only term-pairs with a common term in the window are considered. At each movement, the newest term is retained, then one of the other terms is selected to be retained while the others are removed. The remaining term-pair is mapped to a feature. Hence, $n - 1$ features are extracted at each movement, greatly reducing the number of features compared with SBPH. Experiments show OSB slightly outperforms SBPH in terms of error rate.

D. Concentration Based Feature Construction Approach

Inspired by human immune system, Tan *et al.* proposed concentration based feature construction (CFC) approach for spam filtering [17], [18]. The basic idea is the concentrations of antibodies can be seen as reflection of antigens, for the concentrations increase when antigens invade into human body. The CFC approach computes "self" and "non-self" concentrations by utilizing "self" and "non-self" gene libraries, transforming an email to a 2-dimensional feature vector. A gene denotes an individual term which occurs in the training set. "Self" and "non-self" gene libraries are constructed by computing the difference of frequencies that a term occurs in spam and ham, referred to as proclivity, and terms with high proclivities are selected. Concentration is defined as the number of distinct terms that occur in both gene library and the current email divided by the total number of distinct terms that occur in the current email. The CFC approach essentially decreases the dimension of feature vectors leading to efficiency improvement of spam filtering, and achieves good performance on classification accuracy.

E. Local-Concentration Based Feature Extraction Approach

Zhu *et al.* proposed local-concentration based feature extraction (LC) approach for spam filtering based on that the local concentrations of antibodies determine whether the corresponding pathogens can be culled from the body [19], which can be seen as an improvement of CFC approach. The LC approach is considered to be able to extract position-correlated information by utilizing a sliding window to divide an email into areas. "Self" and "non-self" concentrations are computed independently on each area and finally combined together to form the feature vector of an email. In addition, term selection strategies are employed in the LC approach, and the LC approach redefines term tendency, which is difference of the probabilities of a term's occurrence given the email's class, for constructing the detector sets instead of using proclivity. Experimental results demonstrate that the LC approach remains the feature of high efficiency of CFC approach and outperforms CFC in terms of both classification accuracy and F_1 measure.

III. MRC Based Feature Construction Approach

A. Motivation

Biological immune system (BIS) is an adaptive distributed system with the capability of discriminating "self" and "non-self" and further protecting the biological system from invasion of pathogens. In the BIS, antibodies, produced by lymphocytes, can detect and destroy pathogens by binding them. Two types of immune response may happen in the BIS. The primary response happens when a pathogen appears for the first time and antibodies with affinity to the pathogen are produced slowly. A secondary response is trigged when the same pathogen appears again and a large amount of antibodies with high affinity to the pathogen are proliferated as a corresponding long-lived B memory cell is created during the primary response. Therefore, concentrations of antibodies with affinity to pathogens increase no matter a primary or secondary response happens. Concentrations of antibodies in local areas can reflect the corresponding pathogens precisely. We propose an MRC approach by mimicking the dynamic process of gradual refinement in locating the pathogens by calculating the local concentrations of antibodies, where emails are transformed into multi-resolution concentration feature vectors with respect to "antibodies".

B. Structure of MRC Model

The structure of MRC model, which depicts the dynamic process of gradual refinement in locating the pathogens by calculating local concentrations of antibodies on smaller and smaller resolutions, is shown in Fig. 1. The purpose of preprocessing step in the model is transforming email

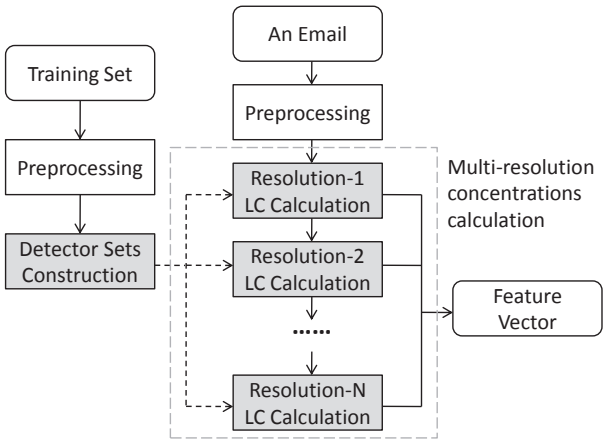

Fig. 1. The structure of MRC model

into terms (words) by examining the existence of blank spaces and delimiters. Detector sets construction and multi-resolution concentrations calculation are essential steps of the MRC model.

1) Detector sets construction: Detector (antibody) sets, utilized for calculating concentrations, are constructed on the training set. A detector represents an individual term. After preprocessing of the training set, huge number of terms are got. In order to reduce the computational complexity of the model and the affect from possible noisy terms, a certain term selection strategy is utilized. For the informative terms reserved, tendency to spam or ham is employed to finish the construction of "Self" detector set and "non-self" detector set.

2) Multi-resolution concentrations calculation: The multi-resolution concentrations calculation step depicts a natural and direct strategy of gradual refinement in locating the pathogens by calculating the local concentrations of detectors. It presents a dynamic process with the resolutions getting smaller and smaller till the concentrations get stable. After finishing the refinement process, all the concentration features calculated are combined together to form the feature vector.

C. MRC Based Feature Construction Approach

Algorithm 1 gives a detailed description of the detector sets construction process, which mainly contains term selection and tendency calculation. The purpose of term selection is to reduce the computational complexity and affect from possible noisy terms. The parameter p determines the extent of term selection.

Information gain (IG) [20], the most frequently employed term goodness criterion in machine learning area, which measures the number of bits of information obtained for text classification by knowing the presence or absence of a term in a text, is selected as the term selection strategy. When applied to spam filtering, information gain of term t_i can be defined as

$$\tau(t_i) = \sum_{c\in(s,h)} \sum_{t\in(t_i,\bar{t}_i)} P(t,c)\log\frac{P(t,c)}{P(t)P(c)} \qquad (1)$$

where c denotes the class of an email, s stands for spam, and h stands for ham. t_i and $\bar{t}_i$ denotes the presence and absence of term t_i respectively.

Algorithm 1 Detector sets construction

```
initialize preselected term set TS_p, "ham" detector set DS_h and "spam" detector set DS_s as empty sets

for each term t_i occurs in the training set do
    calculate evaluation τ(t_i) according to a certain term selection strategy
end for
sort the terms in descending order of evaluation
add the front p% terms to TS_p

for each term t_i in TS_p do
    if tendency(t_i) > 0 then
        add t_i to DS_h
    else
        if tendency(t_i) < 0 then
            add t_i to DS_s
        end if
    end if
end for
```

The tendency of term t_i is defined as

$$tendency(t_i) = P(t_i|c_h) - P(t_i|c_s) \qquad (2)$$

where $P(t_i|c_h)$ is the probability of t_i's occurrence, given the email is ham, and $P(t_i|c_s)$ is the probability of t_i's occurrence, given the email is spam. The difference $tendency(t_i)$ shows which class of emails term t_i tends to occur in. Term t_i with $tendency(t_i) < 0$, which means t_i occurs more frequently in spam than in ham, is added into the spam detector set, and vice versa.

The feature vector is constructed by calculating local concentrations of detectors on smaller and smaller resolutions. Algorithm 2 shows the process of multi-resolution concentrations calculation. For efficiency consideration, the resolution set is initialized to determine the whole process of gradual refinement in locating the pathogens, which is quite essential to the MRC approach. It is initialized as $RS = \{1, 2, 2^2, ..., 2^{n-1}\}$ by adopting the bisection method. Each member in the resolution set denotes a certain resolution, which means how many local areas the given email should be partitioned into. The parameter n determines when the process of gradual refinement should stop.

Concentrations of spam detectors and ham detectors are calculated on each local area achieved under a certain resolution. Spam detectors concentration is defined as

$$SC = \frac{N_s}{N_t} \qquad (3)$$

Algorithm 2 Multi-resolution concentrations calculation

initialize resolution set RS
for each resolution r_i in RS **do**
  partition the given email into local areas according to resolution r_i
  for each local area la_j with respect to resolution r_i **do**
    calculate spam detectors concentration SC_{ij}
    calculate ham detectors concentration HC_{ij}
  end for
end for
combine the achieved concentrations together to form the feature vector

where N_s is the number of distinct terms in the local area which have been matched by detectors in DS_s, and N_t is the number of distinct terms in the local area. Ham detectors concentration is defined similarly, which is

$$HC = \frac{N_h}{N_t} \tag{4}$$

where N_h is the number of distinct terms in the local area which have been matched by detectors in DS_h. Finally, the MRC based feature vector of the given email can be acquired by combining all the achieved concentration feature during the refinement process.

D. WMRC Based Feature Construction Approach

Antibodies play core roles in the BIS. On the surface of antibodies, there are specific receptors which can bind corresponding specific pathogens. All the time, a wide variety of antibodies are circulating in the BIS to detect and destroy different kinds of antigens. Since the invading frequency of different antigens varies, the corresponding specific antibodies have different activity levels, which means some kinds of antibodies are activated more frequently than others. Taking inspiration from this, we propose a WMRC approach by considering the different activity levels of detectors during the refinement process in locating the pathogens.

In the WMRC approach, each detector in the constructed detector sets is given a weight, which depicts the activity level of the detector. The weight is defined as

$$w_i = \frac{\tau(t_i)}{\max\limits_{t_j \in DS_s \cup DS_h} \tau(t_j)} \tag{5}$$

where $\tau(t_i)$ is the evaluation of t_i achieved in the detector sets construction process, and $w_i \in [0, 1]$. The weights of detectors are reflected in the multi-resolution concentrations calculation process. The weighted concentration of spam detectors is defined as

$$SC_w = \frac{\sum\limits_{t_i \in TS_s} w_i}{N_t} \tag{6}$$

TABLE I
EXPRESSIONS OF EVALUATION CRITERIA

Criterion	Expression
Spam Recall	$R_s = \frac{n_{s,s}}{n_{s,s}+n_{s,h}}$
Spam Precision	$P_s = \frac{n_{s,s}}{n_{s,s}+n_{h,s}}$
Accuracy	$A = \frac{n_{s,s}+n_{h,h}}{n_s+n_h}$
F_β	$F_\beta = (1+\beta^2)\frac{R_s P_s}{\beta^2 P_s + R_s}$

where TS_s is the term set with distinct terms in the local area which have been matched by detectors in DS_s, and w_i is the weight of the corresponding detector in DS_s. Similarly, the weighted concentration of ham detectors is defined as

$$HC_w = \frac{\sum\limits_{t_i \in TS_h} w_i}{N_t} \tag{7}$$

where TS_h is the term set with distinct terms in the local area which have been matched by detectors in DS_h.

IV. EXPERIMENTS

A. Corpora

Experiments were conducted on PU1, PU2, PU3, PUA [21] and Enron-Spam [22], which are all benchmark corpora widely used for effectiveness evaluation in spam filtering. PU1 contains 1099 emails, 481 of which are spam. PU2 contains 721 emails, and 142 of them are spam. 4139 emails are included in PU3 and 1826 of them are spam. 1142 emails are included in PUA and 572 of them are spam. Enron-Spam contains 33716 emails, 17171 of which are spam. Emails in the five corpora all have been preprocessed by removing header fields, attachment and HTML tags, leaving subject and body text only. For privacy protection, emails in PU corpora have been encrypted by replacing meaningful terms with specific numbers.

B. Evaluation Criteria

Spam recall, spam precision, accuracy and F_β measure [13] are commonly used evaluation criteria in spam filtering. Spam recall measures the percentage of spam that can be correctly classified, while spam precision measures the percentage of real spam in the emails that are classified as spam. Accuracy measures the percentage of emails that are correctly classified, reflecting the overall performance of spam filtering. And F_β measure is a combination of recall and precision, which reflects the overall performance of spam filtering in another aspect. Usually, $\beta = 1$, and F_1 measure is adopted.

The expression of each evaluation criterion is show in Table I, where $n_{s,s}$ is the number of spam classified as spam, and $n_{s,h}$ is the number of spam classified as ham, $n_{h,s}$ and $n_{l,l}$ are defined similarly. n_l and n_s are the total number of legitimate emails and spam, respectively.

In our experiments, accuracy and F_1 measure are the main evaluation criteria, for they can reflect the overall performance of spam filtering, and precision and recall can be reflected by F_1 measure as well.

C. Experimental Setup

All the experiments were conducted on a PC with E4500 CPU and 2G RAM. 10-fold cross validation is used on PU corpora and 6-fold cross validation is used on Enron-Spam according to the number of parts each of the corpora has been already divided into. WEKA toolkit [23] was utilized in implementation of classification methods, as well as the library LIBSVM [24], which is an implementation of SVM.

D. Investigation of Parameters

Experiments have been conducted on PU3 corpus with SVM as the classification method to investigate the parameters of the MRC approach. 10-fold cross validation was utilized. There are two important parameters as mentioned above. Parameter p determines the percentage of terms reserved after term selection in the detector sets construction process. The removal of less informative terms can reduce not only the computational complexity but also the affect from possible noisy terms, so as to improve the efficiency and performance. While parameter n, which is much more essential, determines the whole process of multi-resolution concentrations calculation, especially when the refinement process should stop.

Fig. 2 describes the performance of the MRC approach under different values of n. As we can see, along with the gradual refinement process, the performance shows improvements till $n = 4$. However, with further increase of n, the performance of the MRC approach degrades in terms of both accuracy and F_1 measure, which reveals the decreased generation capability due to overfitting on training set. Thus, $n = 4$ properly defines the termination of the gradual refinement process.

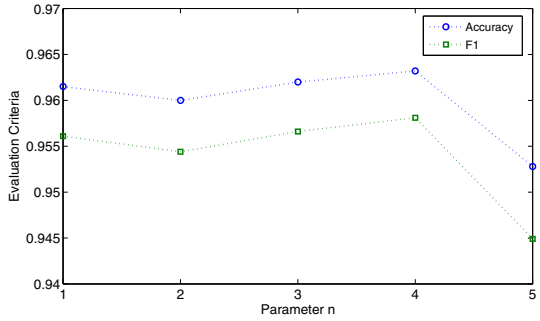

Fig. 2. Performance of the MRC approach with varied n

The performance of the MRC approach with varied p is shown in Fig. 3. As we see, the MRC approach always performs quite well though p varies, but better performance can be achieved with smaller ps, which is consistent with our expectation. The MRC approach performs best with $p = 20$. In this case, the front 20% terms are reserved after the term selection phase.

Similarly, we conducted experiments on PU3 corpus with SVM as the classification method to tune the parameters of the WMRC approach. As the experimental results shown,

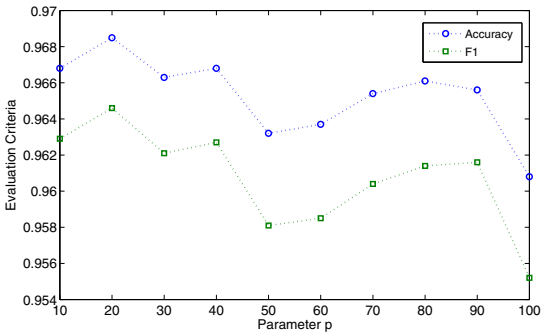

Fig. 3. Performance of the MRC approach with varied p

$n = 3$ and $p = 20$ lead to the best performance, see Fig. 4 and Fig. 5.

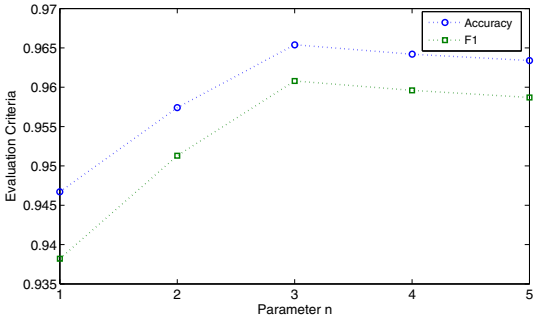

Fig. 4. Performance of the WMRC approach with varied n

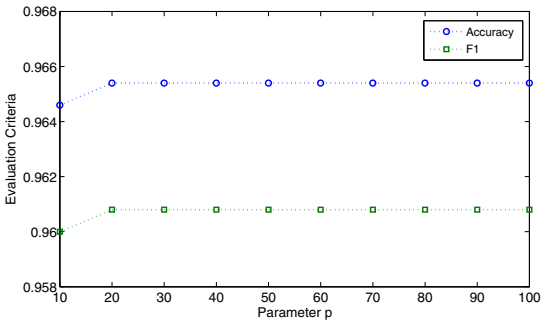

Fig. 5. Performance of the WMRC approach with varied n

E. Comparison with the prevalent approaches

Experiments were conducted on PU1, PU2, PU3, PUA and Enron-Spam to compare the performance of the proposed MRC and WMRC approaches with the prevalent approaches. The selected prevalent approaches are BoW, CFC and LC. Table II shows the spam filtering performance of each feature construction approach when incorporated with SVM, and the corresponding dimensions of feature vectors constructed. As mentioned above, we take accuracy and F_1 measure as comparison criteria without focusing on precision and recall,

TABLE II
PERFORMANCE COMPARISON OF THE MRC AND WMRC APPROACHES WITH THE PREVALENT APPROACHES

Corpus	Approach	Precision(%)	Recall(%)	Accuracy(%)	F_1(%)	Feature dim.
PU1	BoW	93.96	95.63	95.32	94.79	600
	CFC	94.97	95.00	95.60	94.99	2
	LC	94.85	95.63	95.78	95.21	6
	MRC	95.49	95.62	96.06	95.51	30
	WMRC	96.34	96.46	**96.79**	**96.33**	14
PU2	BoW	88.71	79.29	93.66	83.74	600
	CFC	95.12	76.43	94.37	84.76	2
	LC	95.74	77.86	94.79	85.16	6
	MRC	94.16	82.14	95.07	86.98	30
	WMRC	94.44	85.71	**96.06**	**89.37**	14
PU3	BoW	96.48	94.67	96.08	95.57	600
	CFC	96.24	94.95	96.05	95.59	2
	LC	96.68	94.34	96.03	95.45	6
	MRC	96.47	96.54	**96.85**	**96.46**	30
	WMRC	96.58	95.66	96.54	96.08	14
PUA	BoW	92.83	93.33	92.89	93.08	600
	CFC	96.03	93.86	94.82	94.93	2
	LC	95.60	94.56	94.91	94.94	6
	MRC	95.00	95.09	94.91	**94.95**	30
	WMRC	96.95	93.16	**95.00**	94.94	14
Enron-Spam	BoW	90.88	98.87	95.13	94.62	600
	CFC	91.48	97.81	95.62	94.39	2
	LC	92.44	97.81	96.02	94.94	6
	MRC	92.74	98.42	96.29	**95.42**	30
	WMRC	92.76	98.08	**96.30**	95.26	14

which are incorporated into the calculation of F_1 measure and can be reflected by F_1 measure.

As we can see, compared with BoW, the proposed MRC and WMRC approaches not only make significant reduction on the feature vector dimension, but also achieve much better performance, which is average 1.58% higher than BoW in terms of accuracy and 2.15% higher in terms of F_1 measure. This demonstrates the MRC and WMRC approaches are effective.

The CFC approach calculates "self" and "non-self" concentrations on the entire email, which is defined as resolution-1 in the proposed MRC and WMRC approaches, to transform an email into a two-dimensional vector. While the LC approach partitions the email on a single resolution to extract position-correlated information by calculating "self" and "non-self" concentrations on each local area. The experimental results show that the MRC and WMRC approaches outperform both CFC and LC in accuracy and F_1 measure, which verified that the proposed MRC model could effectively extract not only position-correlated information but also process-correlated information through the dynamically gradual refinement process in locating pathogens.

The comparison between MRC and WMRC is to verify whether considering the activity levels of detectors in calculation of concentrations is effective. The results indicate that the WMRC approach performs better than MRC on most cases. Moreover, the introduction of weights enables the WMRC approach reduce the number of resolutions that the email should be partitioned on by accelerating the process of gradual refinement and further reduce the dimension of feature vectors constructed.

We conducted experiments on PU1 with SVM as the classification method to compare the efficiency of the feature construction approaches above. The 10-fold cross validation was utilized. The average speed of processing one email is calculated, as shown in Table III. As we can see, all of the CFC, LC, MRC and WMRC approaches perform far more efficient than BoW, due to significant reduction on feature vector dimension. Although the MRC and WMRC approaches depict the dynamic process of gradual refinement in locating pathogens and increase the feature vector dimension, the processing efficiency decreases not so much. Because less terms are reserved after term selection in MRC and WMRC, and the resolution set is initialized to determine the whole process of multi-resolution concentrations calculation.

F. Performance with other classification methods

To filter spam effectively, both the feature construction approach and the classification method are essential. Since the performance of a feature construction approach must be reflected by cooperating with certain classification methods, it is necessary to verify whether the proposed MRC and

TABLE III

EFFICIENCY COMPARISON OF THE MRC AND WMRC APPROACHES WITH THE PREVALENT APPROACHES

Approach	BoW	CFC	LC	MRC	WMRC
Seconds/email	$9.57e^{-3}$	$3.75e^{-4}$	$4.50e^{-4}$	$6.46e^{-4}$	$5.22e^{-4}$

TABLE IV

PERFORMANCE OF THE MRC APPROACH INCORPORATED WITH DIFFERENT CLASSIFICATION METHODS

Corpus	Classifier	Precision(%)	Recall(%)	Accuracy(%)	F_1(%)
PU1	Naive Bayes	94.92	92.25	96.06	95.54
	C4.5	94.15	93.33	94.50	93.70
	AdaBoost1	94.52	95.21	95.41	94.81
	AdaBoost2	95.50	95.21	95.87	95.29
PU2	Naive Bayes	85.22	87.86	93.80	85.62
	C4.5	89.08	80.00	93.80	83.70
	AdaBoost1	91.97	82.86	94.93	86.57
	AdaBoost2	93.06	79.29	94.37	84.95
PU3	Naive Bayes	95.73	94.07	95.47	94.85
	C4.5	93.89	96.04	95.40	94.88
	AdaBoost1	93.37	96.59	95.38	94.89
	AdaBoost2	95.56	96.26	96.30	95.86
PUA	Naive Bayes	94.57	94.74	94.39	94.50
	C4.5	91.90	94.04	92.63	92.82
	AdaBoost1	92.11	94.21	92.89	93.04
	AdaBoost2	94.96	93.86	94.30	94.30
Enron-Spam	Naive Bayes	90.98	98.68	95.89	94.55
	C4.5	92.43	98.07	95.87	95.07
	AdaBoost1	90.28	98.81	95.48	94.19
	AdaBoost2	93.64	98.50	96.71	95.94

WMRC approaches can be incorporated with different classification methods.

We conducted experiments on PU1, PU2, PU3, PUA and Enron-Spam to investigate the performance of the proposed approaches with different classification methods. The selected classification methods are naive bayes, C4.5 decision tree, AdaBoost with Id3 as the base classifier (AdaBoost1) and AdaBoost with C4.5 as the classifier (AdaBoost2), which are all conmmonly used classification methods in machine learning area. The performance of the MRC approach and the WMRC approach with different classifiers are list in Table IV and Table V, respectively. As we can see, both the MRC and WMRC approaches can perform well with variety of classifiers, which endows them with flexible capability in the real world.

V. CONCLUSIONS

In this paper, we proposed a MRC based feature construction approach for spam filtering by taking inspiration from BIS. Feature construction is considered as a process of gradual refinement in locating the pathogens by dynamically calculating local concentrations of detectors on smaller and smaller resolutions. By introducing activity level of detector, a WMRC based feature construction approach is presented. Sufficient experiments illustrate that the MRC and WMRC approaches outperform prevalent feature construction approaches in spam filtering and achieve high efficiency.

In future work, we intend to incorporate other term selection strategies into the MRC model. In addition, we hope to design a dynamic implementation, which can adaptively adjust the resolutions, of the MRC model.

REFERENCES

[1] L. Cranor and B. LaMacchia, "Spam!" *Communications of the ACM*, vol. 41, no. 8, pp. 74–83, 1998.

[2] Symantec, "Symantec intelligence report: January 2012," *Tech. rep.*, 2012.

[3] Commtouch, "Internet threats trend report-april 2012," *Tech. rep.*, 2012.

[4] F. Research, "Spam, spammers, and spam control: A white paper by ferris research," *Tech. rep.*, 2009.

[5] M. Sahami, S. Dumais, D. Heckerman, and E. Horvitz, "A bayesian approach to filtering junk e-mail," in *Learning for Text Categorization: Papers from the 1998 workshop*, vol. 62. Madison, Wisconsin: AAAI Technical Report WS-98-05, 1998, pp. 98–105.

[6] A. Ciltik and T. Gungor, "Time-efficient spam e-mail filtering using n-gram models," *Pattern Recognition Letters*, vol. 29, no. 1, pp. 19–33, 2008.

[7] H. Drucker, D. Wu, and V. Vapnik, "Support vector machines for spam categorization," *Neural Networks, IEEE Transactions on*, vol. 10, no. 5, pp. 1048–1054, 1999.

[8] J. Clark, I. Koprinska, and J. Poon, "A neural network based approach to automated e-mail classification," in *Web Intelligence, 2003. WI 2003. Proceedings. IEEE/WIC International Conference on.* IEEE, 2003, pp. 702–705.

TABLE V

PERFORMANCE OF THE WMRC APPROACH INCORPORATED WITH DIFFERENT CLASSIFICATION METHODS

Corpus	Classifier	Precision(%)	Recall(%)	Accuracy(%)	F_1(%)
PU1	Naive Bayes	95.96	97.71	97.16	96.80
	C4.5	94.27	95.83	95.60	95.02
	AdaBoost1	96.16	95.62	96.33	95.77
	AdaBoost2	95.88	96.46	96.61	96.13
PU2	Naive Bayes	82.10	93.57	94.08	86.74
	C4.5	88.22	85.00	94.65	86.17
	AdaBoost1	86.20	88.57	94.51	86.69
	AdaBoost2	88.39	87.14	94.93	87.40
PU3	Naive Bayes	95.37	96.26	96.22	95.76
	C4.5	95.21	94.73	95.50	94.92
	AdaBoost1	93.24	95.55	94.87	94.32
	AdaBoost2	95.59	95.99	96.22	95.75
PUA	Naive Bayes	93.92	94.91	94.04	94.25
	C4.5	92.92	94.04	93.07	93.31
	AdaBoost1	91.18	95.79	92.98	93.30
	AdaBoost2	94.30	94.74	94.30	94.40
Enron-Spam	Naive Bayes	91.37	98.48	95.95	94.67
	C4.5	92.59	98.04	96.14	95.13
	AdaBoost1	90.08	98.78	95.37	94.06
	AdaBoost2	93.79	98.46	96.78	95.99

[9] C. Wu, "Behavior-based spam detection using a hybrid method of rule-based techniques and neural networks," *Expert Systems with Applications*, vol. 36, no. 3, pp. 4321–4330, 2009.

[10] I. Androutsopoulos, G. Paliouras, V. Karkaletsis, G. Sakkis, C. Spyropoulos, and P. Stamatopoulos, "Learning to filter spam e-mail: A comparison of a naive bayesian and a memory-based approach," *Arxiv preprint cs/0009009*, 2000.

[11] G. Sakkis, I. Androutsopoulos, G. Paliouras, V. Karkaletsis, C. Spyropoulos, and P. Stamatopoulos, "A memory-based approach to anti-spam filtering for mailing lists," *Information Retrieval*, vol. 6, no. 1, pp. 49–73, 2003.

[12] X. Carreras and L. Marquez, "Boosting trees for anti-spam email filtering," *Arxiv preprint cs/0109015*, 2001.

[13] T. Guzella and W. Caminhas, "A review of machine learning approaches to spam filtering," *Expert Systems with Applications*, vol. 36, no. 7, pp. 10 206–10 222, 2009.

[14] K. Schneider, "A comparison of event models for naive bayes anti-spam e-mail filtering," in *Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics-Volume 1*. Association for Computational Linguistics, 2003, pp. 307–314.

[15] W. Yerazunis, "Sparse binary polynomial hashing and the crm114 discriminator," *the Web.[Online]. Available: http://crm114. sourceforge. net/CRM114 paper. html*, 2003.

[16] C. Siefkes, F. Assis, S. Chhabra, and W. Yerazunis, "Combining winnow and orthogonal sparse bigrams for incremental spam filtering," *Knowledge Discovery in Databases: PKDD 2004*, pp. 410–421, 2004.

[17] Y. Tan, C. Deng, and G. Ruan, "Concentration based feature construction approach for spam detection," in *Neural Networks, 2009. IJCNN 2009. International Joint Conference on*. IEEE, 2009, pp. 3088–3093.

[18] G. Ruan and Y. Tan, "A three-layer back-propagation neural network for spam detection using artificial immune concentration," *Soft Computing-A Fusion of Foundations, Methodologies and Applications*, vol. 14, no. 2, pp. 139–150, 2010.

[19] Y. Zhu and Y. Tan, "A local-concentration-based feature extraction approach for spam filtering," *Information Forensics and Security, IEEE Transactions on*, vol. 6, no. 2, pp. 486–497, 2011.

[20] Y. Yang and J. Pedersen, "A comparative study on feature selection in text categorization," in *MACHINE LEARNING-INTERNATIONAL WORKSHOP THEN CONFERENCE-*. MORGAN KAUFMANN PUBLISHERS, INC., 1997, pp. 412–420.

[21] I. Androutsopoulos, G. Paliouras, and E. Michelakis, *Learning to filter unsolicited commercial e-mail.* " DEMOKRITOS", National Center for Scientific Research, 2004.

[22] V. Metsis, I. Androutsopoulos, and G. Paliouras, "Spam filtering with naive bayes-which naive bayes," in *Third conference on email and anti-spam (CEAS)*, vol. 17, 2006, pp. 28–69.

[23] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. Witten, "The weka data mining software: an update," *ACM SIGKDD Explorations Newsletter*, vol. 11, no. 1, pp. 10–18, 2009.

[24] C. Chang and C. Lin, "Libsvm: a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 2, no. 3, p. 27, 2011.