A NONNEGATIVE SPARSITY INDUCED SIMILARITY MEASURE WITH APPLICATION TO CLUSTER ANALYSIS OF SPAM IMAGES

Yan Gao and Alok Choudhary

Department of EECS, Northwestern University
{ygao@cs, choudhar@ece}.northwestern.edu

Gang Hua

Nokia Research Center, Hollywood
ganghua@gmail.com

ABSTRACT

Image spam is an email spam that embeds text content into graphical images to bypass traditional spam filters. The majority of previous approaches focus on filtering image spam from client side. To effectively detect the attack activities of the spammers and fast trace back the spam sources, it is also essential to employ cluster analysis to comprehensively filter the image emails on the server side. In this paper, we present a nonnegative sparsity induced similarity measure for cluster analysis of spam images. This similarity measure is based on an assumption that a spam image should be represented well by the nonnegative linear combination of a small number of spam images in the same cluster. It is due to the observation that spammers generate large number of varieties from a single image source with different image processing and manipulation techniques. Experiments on a spam image dataset collected from our department email server demonstrated the advantages of the proposed approach.

Index Terms— Nonnegative sparse representation, Image spam filtering, Cluster analysis

1. INTRODUCTION

The success of text document classification techniques on email spam detection [1, 2, 3] has driven spammers to explore new variations of spam emails, among which image spam email has become a new popular weapon. To generate image spam emails, the spammers embed the text content into an image, on which they impose various image processing techniques such as those utilized in CAPTCHA (Completely Automated Public Turing Test to Telll Computers and Humans Apart). Although large amount of end users receive different image spams, these images are substantially visual variations from a small number of spam image sources. By appending texts containing randomly generated words based on normal statistics with the image spam, the spam images can successfully bypass text based spam filters.

Some early work such as SpamAssassin [4] have tried to pull out the embedded texts in the spam images by using OCR (Optical Character Recognition), and then apply text based spam filtering techniques. However, highly accurate OCR may be by itself a more difficult problem than spam image classification, especially when the spammers are performing adversarial manipulation of the image content. This is probably the reason why many recent work has been focusing on directly classifying email image attachments as either spam or non-spam, such as the different image spam hunters [5, 6] and fast image spam classifiers [7]. A supervised or semi-supervised learning machinery is usually leveraged in these image spam classifiers.

Not withstanding their demonstrated success, these direct classification schemes focus on classifying each individual image attachment. It lacks, however, more global analysis of the corpus of image attachments on the sever. Unsupervised clustering analysis of the image corpus may provide more information on the source of spam images. For example, if all the email users on this server received image attachments from the same cluster, then it is highly likely that they are spam images. Further analysis can then be performed to identify the source senders and block them in the future.

Nevertheless, to effectively cluster images, it is essential to have a good visual similarity measure for different images. Previous work has designed different image signatures from diverse image features to define either L1 or L2 norm, weighted or un-weighted, in the feature space as the similarity measures [8, 9]. However, they are not able to adapt to the manifold structure of the image features, as pointed out by Cheng et al [10].

We propose a *nonnegative sparsity induced similarity measure* and apply it for the task of cluster analysis of spam images. The basic proposition we make is that an image should be able to be effectively reconstructed by a small number of other images from the same cluster. We design a quadratic program to calculate such nonnegative sparse representation and a similarity measure is further derived from such a representation. Our experimental results further validate the efficacy of the proposed nonnegative sparsity induced similarity measure for clustering analysis.

In Sec. 2, we will present the system flow chart of an unsupervised image spam detection system by performing cluster analysis. Then we discuss the proposed nonnegative sparsity induced similarity measure in Sec. 3. Experiments are discussed and analyzed in Sec. 4. We finally conclude in Sec. 5.

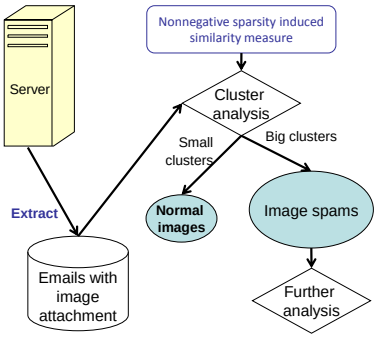

Fig. 1. Flow chart of a server side image spam detection system by cluster analysis.

2. UNSUPERVISED IMAGE SPAM DETECTION

Figure 1 presents the system flow chart of a server side image spam detection system by cluster analysis. Given a set of image attachments extracted from the email server, we cluster them by leveraging the nonnegative sparsity induced similarity measure, which we shall discuss in more detail in Section 3. Since spam images are usually send in bulk, bigger clusters are highly likely to be spam images. They are then sent to either the administrator or an automatic program for further analysis. For example, we can further identify the spam sources so that we can block them from the server side in a very early stage. We shall remark here that this is hard to achieve if we only do client-side spam filtering.

With server side blocking, we hope that the spam emails received by end client users be minimized. Those smaller clusters are most often normal images and so that they will be passed to the client users in the end. There may be false negatives, but they are in small bulk and less annoying to the end users. Moreover, the client side spam image filters could be able to further capture them.

3. NONNEGATIVE SPARSITY INDUCED SIMILARITY MEASURE FOR CLUSTERING

Assume $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N]$ is the feature vectors of the N images we obtained from a batch of emails in an email server, where $\forall i, \mathbf{x}_i \in \mathbf{R}^n$. Our nonnegative sparsity induced similarity is based on a basic assumption. That is, any data sample or feature vector in the corpus can be well represented by the nonnegative linear combination of a small number samples from the same cluster. Nevertheless, for $\mathbf{x}_i$, we do not know beforehand which samples are in the same cluster, not to mention which small set of samples would reconstruct it well.

To successfully identify the potential small sample set to reconstruct $\mathbf{x_i}$, let $\mathbf{X}_i = [\mathbf{x}_1, \ldots, \mathbf{x}_{i-1}, \mathbf{x}_{i+1}, \ldots, \mathbf{x}_N]$, we propose to solve the following optimization problems,

$$\min_{\mathbf{a}} \quad \frac{1}{2}\|\mathbf{x}_i - \mathbf{X}_i \cdot \mathbf{a}\|^2 + \frac{\beta}{2}\|\mathbf{a}\|^2 + \lambda \sum_{j=1}^{n} a_j \qquad (1)$$

$$s.t. \quad \forall j = 1 \ldots N, a_i > 0 \qquad (2)$$

where $\mathbf{a} = [a_1, \ldots, a_{i-1}, a_{i+1}, \ldots, a_N]^T$, and β is a small constant to weight the ridge regression cost to penalize $\mathbf{a}$ with large L2 norm[1]. Since we constrain each a_i to be nonnegative, $f(\mathbf{a}) = \sum_{j=1}^{n} a_j$ is equivalent to an L1 norm Lasso penalty [11]. Therefore, solving the above constrained optimization problem would naturally result in $\mathbf{a}$ to be a sparse vector, i.e., a vector with only a small number of non-zero entries. λ is the control parameter of the Lasso penalty, which directly determines how sparse $\mathbf{a}$ will be.

After easy mathematic derivation, it is straightforward to observe that the above formulation Equation 1 can be re-arranged as

$$\min_{\mathbf{a}} \quad \mathbf{a}^T(\mathbf{X}^T\mathbf{X} + \beta\mathbf{I})\mathbf{a} + (\lambda\mathbf{1} - \mathbf{x}_i^T\mathbf{X})^T\mathbf{a} \qquad (3)$$

$$s.t. \quad \forall j = 1 \ldots N, a_i > 0, \qquad (4)$$

where $\mathbf{I}$ is the identity matrix and $\mathbf{1}$ is a vector with all elements being 1. This is a standard quadratic program with linear constraints and can be solved by standard active set method. We employ the MINQ [12] Matlab library in our implementation to solve it. Notice that the difference of our formulation compared with those of Benaroya [13] is the additional ridge regression term, which is to regularize the solution of linear regression to be more stable.

Naturally, after we have identified the sparse vector $\mathbf{a}$, we define the similarity of $\mathbf{x}_i$ to all the other data samples to be

$$w_{ij} = \frac{a_j}{\sum_{k=1,k\neq i}^{N} a_k}. \qquad (5)$$

Since the w_{ij} induced above may not be symmetric, i.e., $w_{ij} \neq w_{ji}$, our final similarity measure s_{ij} forces it to be symmetric by setting $s_{ij} = \frac{w_{ij}+w_{ji}}{2}$. After we have successfully identified the similarity matrix $S = [s_{ij}]$, we may run any spectral clustering algorithm [14] or a simple hierarchical agglomerative clustering algorithm to clustering the data.

We remark here that our nonnegative sparsity induced similarity measure is partly motivated by the work of Cheng et al. [10]. The most obvious difference is that we introduce the nonnegative constraints into the formulation, while their formulation allows the reconstruction coefficients to be negative, which may not be desirable since it is in conflict with one of the two assumptions the authors made, i.e., a sample can be linearly reconstructed from a small set of samples from the same cluster. The reason is that the negative coefficients have to be forcefully set to zero in their algorithm when defining the final distance measure. Similar to [10], one may also

[1]We fix $\beta = 0.01$ in our experiments.

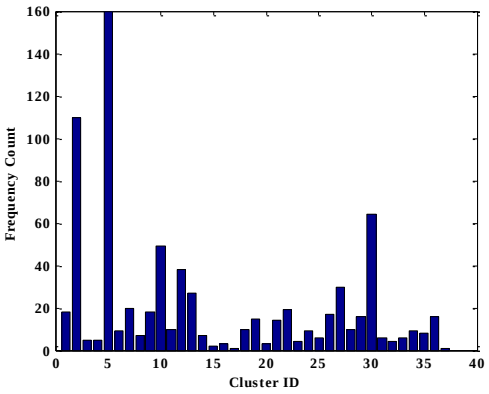

Fig. 2. The number of images in each of the 37 clusters.

pick up the k nearest neighbors of $\mathbf{x}_i$ to form $\mathbf{X}_i$ instead of using all the other $n-1$ data samples, to save the expensive computational cost.[2]

4. EXPERIMENTS

4.1. Data Set

In our previous work [6], we collected an image dataset which contains 1190 spam images and 1760 normal images. Please refer to [6] for details on how we collected the dataset. Among the 1190 spam images, we labeled 37 clusters which covers 756 of the spam images. The number of images in a cluster could be as high as 160, and as low as just 1, as shown in Figure2. These 37 clusters of spam images composed the evaluation data set in our experiments.

4.2. Feature Representation

Following Gao et al [6], we adopt an effective set of 23 image statistics to be the visual representation for each image. It includes 16 color statistic features, 1 texture statistics, 4 shape statistics, and 2 appearance statistics. Due to page limite, we refer to [6] for details of these statistic features.

4.3. Cluster Analysis

We compare the proposed similarity measure with two other competitive measures. The first one is the sparsity induced similarity measure without posing the nonnegative constraint, i.e., we simply remove the nonnegative constraints in Equation 2, set $\beta = 0$ and change $\sum_j a_j$ to $\sum_j |a_j|$ in Equation 1. Then the problem becomes a standard Lasso regression problem [3]. This similarity measure is firstly proposed in [10][4]. The other one is a baseline similarity measure which is induced from the Euclidean distance by applying a Gaussian radial basis function (RBF). For each of the similarity measures, we build the similarity graph matrix and use a spectral cluster algorithm [14] to generate the clustering results.

4.3.1. Evaluation Criterion

Since we have the ground truth cluster labels of all the data, we use two criteria to evaluate the performance of all the clustering results [15]. The first criterion is the average clustering accuracy CAC, which is defined as

$$CAC = \frac{1}{n}\sum_{i=1}^{n}\delta(m(cl_i) - l_i) \tag{6}$$

where n is the total number of data points, $\delta(.)$ is the Dirac-Delta function, cl_i and l_i are the cluster id and the labeled cluster id of data point i, respectively, and $map(\cdot)$ is the best map of cl_i to the ground truth cluster id, which can be optimally resolved by the Kuhn-Munkres algorithm.

The second evaluation criterion we adopt is the normalized mutual information [15] between the cluster results C' and the ground truth clusters C, which is defined as

$$\mu MI = \frac{MI(C, C')}{\max(H(C), H(C'))} \tag{7}$$

where $H(\cdot)$ represents the entropy of the cluster set and $MI(C, C')$ is the mutual information between the two cluster sets, i.e.

$$MI(C, C') = \sum_{c_i \in C, c'_j \in C'} p(c_i, c'_j) \log \frac{p(c_i, c'_j)}{p(c_i)p(c'_j)}. \tag{8}$$

It is easy to figure out that $\mu MI \in [0, 1]$, with $\mu MI = 0$ if the two cluster sets are independent and $\mu MI = 1$ if the two cluster sets are identical.

4.4. Comparison Results

We summarize the cluster performance using three different distance measures in Table 1. We name the results of three different similarity measures as NonNegSparse (our proposed one), Sparse (Cheng et al. [10]), and Euclidean (Gaussian RBF baseline). Since the final step of the spectral clustering algorithm [14] is running a k-means, each run of the spectral cluster will result in slightly different clustering results due to different initialization of the k-means iterations. Therefore, we run the spectral clustering 500 times for each case and the results reported in the table are the mean value plus/minus the standard deviation over all the runs.

As we can clearly observe, the proposed nonnegative sparsity induced similarity measure achieves the best clustering performance with $CAC = 0.635$ and $\mu MI = 0.734$, with a parameter setting $\lambda = 0.1$. This significantly improves the best results achieved by Cheng et al. [10], which obtains $CAC = 0.559$ and $\mu MI = 0.671$ with $\lambda = 0.7$. Nevertheless, both algorithms lead the baseline Gaussian RBF similarity by a significant margin ($CAC = 0.485$ and $\mu MI =$

[2] We fix $k = 100$ in our experiments.

[3] We solve it with Gaussian-Seidel method using the Matlab code provided at http://people.cs.ubc.ca/ schmidtm/Software/lasso.html

[4] Cheng et al. [10] cast it in a slightly different optimization problem, but it should essentially achieve very similar results.

	NonNegSparse	Sparse	Euclidean
CAC	0.635± 0.006	0.559± 0.032	0.485± 0.019
μMI	0.734± 0.005	0.671± 0.006	0.471± 0.025
	$\lambda = 0.1$	$\lambda = 0.7$	–

Table 1. The clustering performance of three different similarity measures.

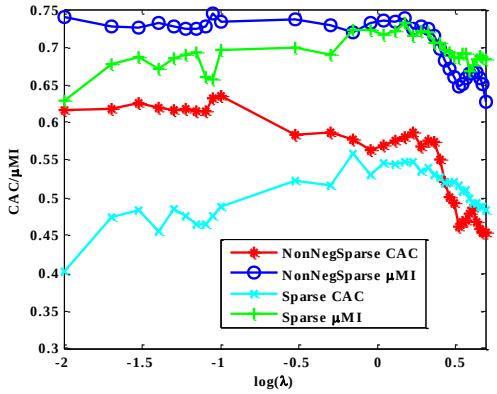

Fig. 3. The changes of cluster performance criteria w.r.t. the λ (in log scale).

0.471), as shown in the table. The standard deviations of the performance quantities of the proposed approach also seem to be smaller than those of the competition methods, which is an indication that the proposed similarity measure is more preferable since the clustering results from it are less sensitive to the initialization of k-means after spectral embedding.

We shall remark here that the weight factor λ in Equation 1 has an impact on both the nonnegative sparsity induced similarity measure and the sparsity induced similarity measure. Therefore, we run the above cluster analysis with different settings of λ for both algorithms, and plot the changes of the cluster performance criteria w.r.t. λ (in log scale) in Figure 3. It clearly demonstrates the better performance of the proposed nonnegative sparsity induced similarity measure.

We notice that the two evaluation criteria, CAC and μMI, are not always strictly tied with one another. That is, when CAC achieves the optimal value, the μMI may not achieve the best simultaneously, and vice versa. We regard CAC as a more direct criterion, so we pick up the working parameter λ based on it in Table 1.

5. CONCLUSION AND FUTURE WORK

In this paper, we present a nonnegative sparsity induced similarity measure and apply it for the task of cluster analysis of spam images by server side. Our experimental comparisons present favorable performance when compared with previous sparsity induced similarity measure without nonnegative constraints and Euclidean distance similarity metric by applying a Gaussian radial basis function. Our future work may include further analysis of the proposed similarity measure in other signal processing tasks.

6. REFERENCES

[1] Mehran Sahami, Susan Dumais, David Heckerman, and Eric Horvitz, "A bayesian approach to filtering junk e-mail," in *Proc. AAAI Workshop on Learning for Text Categorization*, Madison, Wisconsin, July 1998.

[2] Harris Drucker, Donghui Wu, , and Vladimir N. Vapnik, "Support vector machines for spam categorization," *IEEE Transactions on Neural Networks*, vol. 10, pp. 1048–1054, 1999.

[3] Xavier Carreras and Jordi Girona Salgado, "Boosting trees for anti-spam email filtering," in *Proc. the 4th International Conference on Recent Advances in Natural Language Processing*, Tzigov Chark, BG, 2001, pp. 58–64.

[4] SpamAssassin, "http://spamassassin.apache.org/," .

[5] Yan Gao, Ming Yang, Xiaonan Zhao, Bryan Pardo, Ying Wu, Thrasyvoulos Pappas, , and Alok Choudhary, "Image spam hunter," in *Proc. of the 33th IEEE International Conference on Acoustics, Speech, and Signal Processing*, Las Vegas, NV, USA, April 2008.

[6] Yan Gao, Ming Yang, and Alok Choudhary, "Semi supervised image spam hunter: A regularized discriminant em approach.," in *Proc. 5th International Conference Advanced Data Mining and Applications*, Beijing, China, August 2009, vol. LNCS 5678, pp. 152–164.

[7] Mark Dredze, Reuven Gevaryahu, and Ari Elias-Bachrach, "Learning fast classifiers for image spam," in *Proc. the 4th Conference on Email and Anti-Spam (CEAS)*, California, USA, August 2007.

[8] Zhe Wang, William Josephson, Qin Lv, Moses Charikar, and Kai Li, "Filtering image spam with near-duplicate detection," in *Proc. the 4th Conference on Email and Anti-Spam (CEAS)*, California, USA, August 2007.

[9] Bhaskar Mehta, Saurabh Nangia, Manish Gupta, and Wolfgang Nejdl, "Detecting image spam using visual features and near duplicate detection," in *Proc. the 17th International World Wide Web Conference*, Beijing, China, April 2008.

[10] Hong Cheng, Zicheng Liu, and Jie Yang, "Sparsity induced similarity measure for label propagation," in *Proc. IEEE International Conf. on Computer Vision*, Kyoto, Japan, October 2009.

[11] Robert Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society, Series B*, vol. 58, pp. 267–288, 1994.

[12] Arnold Neumaier, "Minq-general definite and bound constrained indefinite quadratic programming," in *http://www.mat.univie.ac.at/˜neum/software/minq/*, 1998.

[13] Laurent Benaroya, Lorcan McDonagh, Frederic Bimbot, and Remi Bribonval, "Non negative sparse representation for wiener based source separation with a single sensor," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, April 2003, vol. 6, pp. 613–616.

[14] Yangqiu Song, Wen-Yen Chen, Hongjie Bai, Chih-Jen Lin, and Edward Y. Chang, "Parallel spectral clustering," in *Proc. European Conference on Machine Learning and Knowledge Discovery in Databases*, Antwerp, Belgium, September 2008.

[15] Deng Cai, Xiaofei He, Wei Vivian Zhang, and Jiawei Han, "Regularized locality preserving indexing via spectral regression," in *Proc. 2007 ACM Int. Conf. on Information and Knowledge Management*, Lisboa, Portugal, November 2007.