Content-Based Spam Filtering Using Hybrid Generative Discriminative Learning of Both Textual and Visual Features

Ola Amayri
Electrical and Computer Engineering Department
Concordia University
Montreal, Canada, H3G 2W1
Email: o_amayri@ece.concordia.ca

Nizar Bouguila
Concordia Institute for Information Systems Engineering
Concordia University
Montreal, Canada, H3G 2W1
Email: bouguila@ciise.concordia.ca

***Abstract*—In this paper, we propose a hybrid generative discriminative framework for the challenging problem of spam emails filtering using both textual and visual features. Our framework is based on building probabilistic Support Vector Machines (SVMs) kernels from mixture of Langevin distributions. Through empirical experiments, we demonstrate the effectiveness and the merits of the proposed learning framework.**

***Keywords*-Spam; SVM; Langevin mixture, discriminative learning; generative learning; probabilistic kernels; bag of words; local features.**

I. Introduction

The diversified spam structures, sophisticated spammer mechanisms, unique features and immense volume of spam emails in unrelated subjects have made the design of anti-spam filters difficult and challenging. The main approach to build automated adaptive classifiers to discriminate legitimate and spam emails has been the content-based analysis of emails. In particular, the analysis of the semantic textual content via the adoption of text categorization techniques, based generally on machine learning and pattern recognition approaches, has received a lot of attention in the past and has achieved acceptable results as compared to techniques based on hand-made rules where emails are described as bags of words (BoW). However, this approach has been defeated in the past by spammers using very simple tricks such as misspelling words or adding bogus text to their emails. Unfortunately, the dynamic nature and the diversity of spam emails have easily circumvented the majority of the existing spam filters especially if we take into account the fact that email's content has massive shift from text content only to enriched multimedia content. Indeed, very recently researchers have figured out that text-based techniques might be ineffective because of a novel spammers trick namely image-based spam (i.e. email which includes embedded image) [1].

Image-based spam email circumvents easily classic text based spam filters, thus some approaches have been proposed to detect the nature of email from its image content. For instance, three plug-ins have been developed within the *SpamAssassin* project to analyze text embedded into images, where they only provide a boolean output that indicates if more than one specific keywords are detected, using an optical character recognition (OCR) system, in the attached image. In the same vein, authors in [2] extract features, related to the presence of text in spam images, which are fed to SVM used as the base classifier. An SVM-based approach which exploits, via OCR tools, the text information embedded into images sent as attachment has been proposed in [1] which is actually the first study that has addressed the problem of analyzing, and not only detecting, the text information embedded into images. All these approaches consider, however, only the textual content of the image and ignore its rich low-level visual content which can be very helpful as previously has proven. Only few papers have considered the low level visual content of spam images as a solution to make filters more robust and smarter. A complementary approach to [1] that aimed at detecting, via two image quality measures, the use of obscuring techniques which can make OCR ineffective instead of extracting textual information from the image was proposed in [3] and equipped the Spam Assassin filter with a new plugin called *Image Cerberus*. Motivated by the recent success of local descriptors in computer vision applications, the authors in [4] proposed the modeling of images using the so-called visual keywords (i.e. quantization of local descriptors) which are then classified as spam or ham using SVM. This approach has some merits since the local descriptor used (i.e. SIFT) is robust to several geometric transformations that may be used by spammers, but the quantization step applied can cause the loss of important information about the image content. It is noteworthy that all these previous approaches were completely text-free.

In this work, we propose a hybrid generative discriminative learning approach that combines and uses simultaneously both textual and visual information to filter spam emails. While the textual content is represented using the classic BOW formalism, the visual image spam information is represented as a bag of local descriptors extracted from detected image keypoints. To the best of our knowledge there is no

978-1-4673-0219-7/12/$31.00 ©2012 IEEE

prior research work in considering simultaneously, within the same statistical framework, textual and low-level visual contents of spams. Moreover, unlike our work, all the previous learning approaches have considered either generative (e.g. GMM) or discriminative models (e.g. SVM, maximum entropy). We are mainly motivated by recent research works that have shown theoretically and experimentally that hybrid models outperform significantly both their discriminative and generative counterparts (see, for instance, [5]). Our hybrid framework is based on finite Langevin mixture [6] as a generative model and SVM as a discriminative classifier. The choice of SVM is motivated by the fact that it has became a standard learning tool leading to benchmark results due to its generalization ability and computational efficiency especially in high-dimensional feature spaces and for the specific problem of spam filtering [7]. The use of SVM is, however, challenging when examples are represented as set of features which may vary in cardinality. Classic kernels (e.g. polynomial) cannot be deployed in this situation which is exactly our case since spam images are represented as sets of local descriptors within our framework. We tackle this problem by modeling these descriptors, in an unsupervised way, using finite mixtures of Langevin distribution (i.e. each spam image is represented by a finite Langevin mixture model) from which probabilistic kernels are generated. This can be viewed as a mapping from a feature space to a probabilistic one where the classification problem is reduced to the comparison of statistical models rather than vectors. It is noteworthy that the deployment of Langevin mixtures allows a natural representation of sparse data in high dimensional spaces and implies the need for a preprocessing stage which consists on L_2 normalizing the feature vectors so they lie on a unit hypersphere. This fact motivates further the consideration of Langevin mixture and SVM within a hybrid framework for the very specific problem of spam filtering, since L_2 normalization has been shown to play an important role, as a preprocessing type, in SVM-based classification[1] especially when dealing with the problem of spam filtering [7].

The rest of this paper is organized as follows. In Section II we briefly introduce the Langevin mixture. In Section III, we derive SVM probabilistic kernels based on Langevin mixture model. We demonstrate the capability and merits of the proposed approach in content-based spam filtering in Section IV. Finally, we sum up and conclude the paper in Section V.

II. Finite Langevin Mixture Model

Let $\vec{X} = (X_1, \ldots, X_p)$ be a random unit vector in $\mathbb{R}^p$. $\vec{X}$ is said to have a p-variate Langevin distribution if its probability density function is given by [6]:

$$p_p(\vec{X}|\vec{\mu}, \kappa) = \frac{\kappa^{\frac{p}{2}-1}}{(2\pi)^{\frac{p}{2}} I_{\frac{p}{2}-1}(\kappa)} \exp\{\kappa \vec{\mu}^T \vec{X}\} \qquad (1)$$

on the $(p-1)$-dimensional unit sphere $\mathbb{S}^{p-1} = \{\vec{X} | \vec{X} \in \mathbb{R}^p : ||\vec{X}|| = (\vec{X}^T \vec{X}) = 1\}$, with mean direction unit vector $\vec{\mu} \in \mathbb{S}^{p-1}$, where $\vec{\mu}^T$ denotes the transpose of $\vec{\mu}$ and non-negative real concentration parameter $\kappa \geq 0$. Furthermore, $I_p(\kappa)$ denotes the modified Bessel function of first kind and order p [6]. From Eq. 1 we can notice that Langevin distribution is a member of (curved)-exponential family of order p, whose shape is symmetric and unimodal, with minimal canonical parameter $\kappa\vec{\mu}$ and minimal canonical statistic $\vec{X}$. Let $p(\vec{X}_i|\Theta)$ be a mixture of M Langevin distributions:

$$p(\vec{X}_i|\Theta) = \sum_{j=1}^{M} p_p(\vec{X}_i|\theta_j) p_j \qquad (2)$$

where $\Theta = \{\vec{P} = (p_1, \ldots, p_M), \vec{\theta} = (\theta_1, \ldots, \theta_M)\}$ denotes all the parameters of the mixture model such as $\theta_j = (\mu_j, \kappa_j)$ and $\vec{P}$ represents the vector of clusters probabilities (i.e. mixing weights) which must be positive and sum to one. The complete learning algorithm for Langevin mixture can be found in [9].

III. Spam Filter Design

A. Emails Presentation

For textual part, we extracted features from the body, subject and header information presented in sender (From, Reply-to) and recipient (To, CC, Bcc) fields. Each email $\vec{X}_i$ was described in terms of counts of features that appear in the dictionary $\vec{X}_i = (X_{i1}, \ldots, X_{ij}, \ldots, X_{i|\mathcal{D}_{Term}|})$, where X_{ij} presents the frequency of the j-th word in the dictionary D_{Term} that appears in the i-th email. In order to resist *sparse data attack* [10] we normalized each feature vector $\vec{X}_i$ using L_2 normalization. Then, we model the probability distribution for each class by Langevin mixture distribution. After we learnt our mixture model, this gives us the text matrix G_{Text} (includes text in the body and in the image). In spam image we classified each image according to its textual and visual content. First, we performed OCR to extract the written part (if presented) from the image and add it to the dictionary of text extracted from email textual parts. Next, to extract visual features, first we use difference-of-Gaussian (DoG) to extract patches around detected interest points. Then, we used SIFT descriptor computed on detected key points of all images resulting in feature vectors with 128 dimensions for each detected interest point in each image. Finally, we normalize extracted vectors using L_2 normalization. Formally, each image is presented now as $I = \{\vec{I}_1, \ldots, \vec{I}_n\}$ where n is the total number of detected key points and $\vec{I}_i$ is the SIFT vector associated with that detected point. Thus, each image in the dataset has been represented

[1] In particular, the authors in [8] recommended strongly the normalization of data in feature space when considering SVM and have shown that normalization leads to considerably superior generalization performance.

by a Langevin mixture which can be viewed actually as our generative step. Then, the Fisher and probabilistic kernels (see section III) between each of these mixture models are computed giving us visual matrix G_{Visual} to feed SVM classifier which represents our discriminative stage.
In an attempt to simultaneously classify emails using their images and text, we combined the resulted text matrix G_{Text} and visual matrix G_{Visual} using α parameter (in the following experiments we set $\alpha = 0.5$ as we suppose that the visual and textual parts have the same importance). Thus, $G_{Total} = \alpha G_{Text} + (1-\alpha)G_{Visual}$.
In the following subsections, we derive different kernels, from Langevin mixture, based on probabilistic distances and Fisher score to tackle the problem of spherical data sequences classification using SVM.

B. Selecting The Kernel

1) Fisher Kernels: Authors in [11] have shown that a generative model can be used in a discriminative context by extracting Fisher scores $U_{\mathcal{X}}(\Theta) = \nabla \log(p(\mathcal{X}|\Theta))$ from the generative model and converting them into a Gram Kernel usable by SVMs. Each component of $U_{\mathcal{X}}(\Theta)$ is the derivative of the log-likelihood of the sequence $\mathcal{X}$ with respect to particular parameter. In the following, we shall show the derivations of the Fisher kernel $\mathcal{K}(\mathcal{X}, \mathcal{X}') = U^{tr}_{\mathcal{X}}(\Theta) I^{-1}(\Theta) U_{\mathcal{X}'}(\Theta')$ for M-Langevin mixture models, where $I(\Theta)$ is the Fisher information matrix. Through the computation of gradient of the log probability with respect to our model parameters, we obtain

$$\frac{\partial \log p(\mathcal{X}|\Theta)}{\partial p_j} = \sum_{i=1}^{N} \Big[\frac{\hat{Z}_{ij}}{p_j} - \frac{\hat{Z}_{i1}}{p_1} \Big] \quad j = 2, \ldots, M \tag{3}$$

$$\frac{\partial \log p(\mathcal{X}|\Theta)}{\partial \vec{\mu}_j} = \sum_{i=1}^{N} \hat{Z}_{ij} \kappa_j \vec{X}_i \tag{4}$$

$$\frac{\partial \log p(\mathcal{X}|\Theta)}{\partial \kappa_j} = \sum_{i=1}^{N} \hat{Z}_{ij} \Big[\vec{\mu}_j^T \vec{X}_i - \acute{a}_p(\kappa_j) \Big] \tag{5}$$

where $\hat{Z}_{ij} = \frac{p(\vec{X}_i|\Theta) p_j}{\sum_{j=1}^{M} p(\vec{X}_i|\Theta) p_j}$ represents the probability that a vector $\vec{X}_i$ will be assigned to cluster j. It is noteworthy that in Eq.3, we take into account the fact that the sum of the mixing parameters equals one and thus there are only $M-1$ free mixing parameters.

2) Probability Product Kernels (PPK): PPK [12], is another approach that maps data points in the input space to distributions over the sample space and a general inner product is then evaluated as the integral of the product of pairs of distributions. let $\mathcal{X} = (\vec{X}_1, \ldots, \vec{X}_N)$ and $\acute{\mathcal{X}} = (\vec{\acute{X}}_1, \ldots, \vec{\acute{X}}_{\acute{N}})$ be two sequences of spherical feature vectors representing two multimedia objects O and $\acute{O}$, respectively, and modeled by two Langevin finite mixtures $p(\vec{X}|\Theta)$ and $q(\vec{X}|\acute{\Theta})$, respectively, defined on Ω space (Ω is the p-dimensional space of Langevin distribution). PPK is defined as

$$\mathcal{K}(p(\vec{X}|\Theta), q(\vec{X}|\acute{\Theta})) = \int_{\Omega} p(\vec{X}|\Theta)^{\rho} q(\vec{X}|\acute{\Theta})^{\rho} d\vec{X} \tag{6}$$

where ρ is a positive parameter. In the case of Langevin distribution, we can find a closed-form expression for the PPK and is given by

$$\int_{\Omega} p(\vec{X}|\theta)^{\rho} q(\vec{X}|\acute{\theta})^{\rho} d\vec{X} = \Big[\Big(\frac{\kappa\acute{\kappa}}{4} \Big)^{\frac{p}{2}-1} \frac{1}{(2\pi)^p I_{\frac{p}{2}-1}(\kappa) I_{\frac{p}{2}-1}(\acute{\kappa})} \Big]^{\rho} \times \Big[\frac{(2\pi)^{\frac{p}{2}} I_{\frac{p}{2}-1}(\xi_{\kappa,\acute{\kappa}}\rho)}{(\xi_{\kappa,\acute{\kappa}}\rho)^{\frac{p}{2}-1}} \Big] \tag{7}$$

where $\xi_{\kappa,\acute{\kappa}} = \sqrt{\kappa^2 + \acute{\kappa}^2 + 2\kappa\acute{\kappa}(\vec{\mu}\vec{\acute{\mu}})}$ and $\tau_{\vec{\mu},\vec{\acute{\mu}}} = \frac{\kappa\vec{\mu} + \acute{\kappa}\vec{\acute{\mu}}}{\xi_{\kappa,\acute{\kappa}}}$ are the concentration parameter and mean direction defining the Langevin distribution, respectively, that results from the multiplication of two Langevin distributions: $p_p(\vec{X}|\vec{\mu},\kappa) p_p(\vec{X}|\vec{\acute{\mu}},\acute{\kappa}) \propto p_p(\vec{X}|\tau_{\vec{\mu},\vec{\acute{\mu}}}, \xi_{\kappa,\acute{\kappa}})$. A special case of PPK is when $\rho = 1$, which is called Expected Likelihood Kernel (ELK). When $\rho = \frac{1}{2}$ PPK has the form of Bhattacharyya kernel (BK) based on Bhattacharyya's measure of affinity between distributions. In the absence of closed form for mixture models, we can approximate PPK using Monte Carlo simulations [12].

IV. Experimental Results

A spam filter has been constructed to validate our proposed spam classification framework, as we consider three main tasks. Firstly, we compare the use of various features described spam email, including text only, visual only, and both simultaneously. Secondly, we trained and tested our dataset with the three kernels we proposed in section III-B which are: Bhattacharyya kernel (BK), Expected likelihood kernel (ELK) and Fisher kernel (FK) and exploring that in terms of Precision (SP), Recall (SR) and Accuracy. Thirdly, in the seek of comparison we compare the performance of hybrid finite multivariate Gaussian mixture with hybrid Langevin mixture models. The libsvm[2] software was used for SVMs classifier. We used stratified 10-fold cross validation to train and test each dataset, and averaged results were reported. The Tesseract OCR suite [3] was used to recognize the embedded text in images.

A. Results

To evaluate the performance of our proposed hybrid framework for spam classification, we used publicly available datasets that have been used in the past. The trec05-p1 dataset[4] where we extracted 1530 images including 1256

[2]http://www.csie.ntu.edu.tw/ cjlin/libsvm/
[3](open source by Google) http://code.google.com/p/tesseract-ocr/
[4]http://plg1.cs.uwaterloo.ca/cgi-bin/cgiwrap/gvcormac/foo

Table I
ACCURACY (IN %± VARIANCE) FOR TEXT, IMAGE, TEXT+ IMAGE BASED SPAM FILTERING BY LANGEVIN AND GAUSSIAN MIXTURES

	Text	Image	Text+Image
Langevin mixture	90.30± 0.4	85.46±1.3	92.59±1.1
Gaussian mixture	86.97±0.91	80.11±2.1	90.82±0.5

Table II
RESULTS (IN%) ON TERC05-P1 USING DIFFERENT TESTED KERNELS FOR LANGEVIN AND GAUSSIAN MIXTURES. ALL RESULTS REPORTED IN THE TABLE ARE FOR "IMAGE +TEXT" AS IT HAS SHOWN TO PROVIDE BEST PERFORMANCE.

Kernels	Langevin Mixture		Gaussian Mixture	
	SP	SR	SP	SR
ELK	75.29	69.15	56.98	66.01
BK	88.59	91.00	85.19	88.90
FK	70.01	79.34	68.74	54.78

spam images and 274 legitimate images. Table I shows the accuracy for spam filtering using Langevin and Gaussian mixtures, respectively, when considering: image, text and image+text based classifiers. Obtained results show that "text+image"-based classifier provides a slight improvement over text-based classifier. However, image-based classifier reported the worst performance, that might be because of the lack of images in our dataset and particularly the images in legitimate emails. Moreover, in all cases results of Langevin mixture are more accurate and precise than those obtained using Gaussian mixture model.

In the next experiment, we fed SVM spam classifier with probabilistic kernels generated from Langevin and Gaussian mixtures (See Table II). In particular, ELK has the worst accuracy among those kernels. This degradation in performance supports the fact that linear kernels are not expressive enough for nonlinear high dimensional spaces. Also, we can see that the use of BK works better than FK, these results support the previous results that FK does not preserve the nonlinearities of given generative model. Moreover, we find that kernels derived based on Langevin is better choice than Gaussian mixture. This can be explained by the fact that the Gaussian mixture, which clustering is based implicitly on the Euclidean distance or Mahalanobis, is inadequate for characterizing L_2 normalized data which clustering structure is better uncovered by considering the cosine similarity as assumed by the Langevin mixture.

V. CONCLUSION

In this paper, we have proposed a spam filtering framework adapted to the enriched multimedia content of emails. The main motivation was the fact that spammers have recently adopted image spam to defeat widely used text categorization based approaches. We have empirically proved the simultaneous exploitation of both textual and visual information achieves better filtering results. The proposed framework is based on a hybrid generative discriminative approach which consists on developing probabilistic SVMs kernels from Langevin mixture. As a typical *adversarial* problem, spam filtering has to be deployed in online settings and hence a possible extension that we are actively working on.

ACKNOWLEDGEMENTS

The completion of this research was made possible thanks to the Natural Sciences and Engineering Research Council of Canada (NSERC).

REFERENCES

[1] G. Fumera, I. Pillai, and F. Roli, "Spam filtering based on the analysis of text information embedded into images," *Journal of Machine Learning Research*, vol. 7, pp. 2699–2720, 2006.

[2] C. T. Wu, K. T. Cheng, Q. Zhu, and Y. L. Wu, "Using visual features for anti-spam filtering," in *Proceedings of IEEE International Conference of Image Processing (ICIP)*, vol. 3, 2005, pp. 501–504.

[3] B. Biggio, G. Fumera, I. Pillai and F. Roli, "Improving image spam filtering using image text features," in *Proceedings of the Conference on Email and Anti-Spam (CEAS)*, 2008.

[4] J.-H. Hsia and M.-S. Chen, "Language-model-based detection cascade for efficient classification of image-based spam e-mail," in *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*, 2009, pp. 1182–1185.

[5] N. Bouguila and O. Amayri, "A discrete mixture-based kernel for svms: Application to spam and image categorization," *Information Processing and Management*, vol. 45, no. 6, pp. 631–642, 2009.

[6] K. V. Mardia, *Statistics of directional data*. Academic Press, 1972.

[7] O. Amayri and N. Bouguila, "A study of spam filtering using support vector machines," *Artif. Intell. Rev.*, vol. 34, no. 1, pp. 73–108, 2010.

[8] R. Herbrich and T. Graepel, "A PAC-Bayesian Margin Bound for Linear Classifiers: Why SVMs Work," in *Proceedings of Advances in Neural Information Processing Systems (NIPS)*, 2000, pp. 224–230.

[9] O. Amayri and N. Bouguila, "Probabilistic clustering based on langevin mixture," in *Proceedings of The 10th International Conference on Machine Learning and Applications (ICMLA'11)*, 2011.

[10] S. Wittel, G.L. Wu, "On attacking statistical spam filters," in *Proceedings of the First Conference on Email and Anti-Spam (CEAS)*, California, USA, 2004.

[11] T. S. Jaakkola and D. Haussler, "Exploiting generative models in discriminative classifiers," in *Proceedings of Advances in Neural Information Systems (NIPS)*. MIT Press, 1998, pp. 487 – 493.

[12] A. B. Chan, N. Vasconcelos, and P. J. Moreno, "A family of probabilistic kernels based on information divergence," University of California, SanDiego, TechnicalReport SVCL-TR2004/01, 2004.