Apply Stacked Auto-Encoder to Spam Detection

Guyue Mi[1,2], Yang Gao[1,2], and Ying Tan[1,2(✉)]

[1] Key Laboratory of Machine Perception (MOE), Peking University, Beijing, China
[2] Department of Machine Intelligence, School of Electronics Engineering and Computer Science, Peking University, Beijing 100871, China
{gymi,gaoyang0115,ytan}@pku.edu.cn

Abstract. In this paper, we apply Stacked Auto-encoder, one of the main types of deep networks, hot topic of machine learning recently, to spam detection and comprehensively compare its performance with other prevalent machine learning techniques those are commonly used in spam filtering. Experiments were conducted on five benchmark corpora, namely PU1, PU2, PU3, PUA and Enron-Spam. Accuracy and F_1 measure are selected as the main criteria in analyzing and discussing the results. Experimental results demonstrate that Stacked Auto-encoder performs better than Naive Bayes, Support Vector Machine, Decision Tree, Boosting, Random Forest and traditional Artificial Neural Network both in accuracy and F_1 measure, which endows deep learning with application in spam filtering in the real world.

Keywords: Spam detection · Machine learning · Artificial neural network · Deep learning · Stacked auto-encoder

1 Introduction

Email has become one of the most commonly used communication tools in our daily work and life due to its advantages of low cost, high efficiency and good convenience. However, the above characteristics are also concerned and exploited by the ones who want to spread advertisement, bad information or even computer virus to send spam emails. Spam, generally defined as unsolicited bulk email (UBE) or unsolicited commercial email (UCE) [1], has caused many problems to our normal email communication. Ferris Research Group [2] has revealed that large amount of spam not only occupied network bandwidth and server storage, but also wasted users' time on reading and deleting them, which resulted in loss of productivity. Moreover, the spam with malware threatens internet safety and personal privacy.

According to Symantec Internet Security Threat Report 2014 [3], although the total number of bots (computers that are infected and controlled to send spam) worldwide has declined from 3.4 million to 2.3 million in 2013 compared with that of 2012, the overall spam rate only dropped 3%, which is still up to 66% of the whole email traffic. What's worse, the phishing rate and virus rate both

Y. Tan et al. (Eds.): ICSI-CCI 2015, Part II, LNCS 9141, pp. 3–15, 2015.
DOI: 10.1007/978-3-319-20472-7_1

increased. In 2013, one out of every 196 emails contained virus and one out of every 392 emails was identified as phishing, while the corresponding proportions in 2012 were 1 in 291 and 1 in 414 respectively. In addition, adult, sex and dating related spam dominated in 2013 and made up 70% of the total spam, which was an increase of 15% compared with that of 2012. The statistics from Cyren Internet Threats Trend Report [4] demonstrate that spam made up 68% of all global emails in the third quarter of 2014, with a daily average of 56 billion. Thus, it is still necessary and urgent to take measures to solve the spam problem.

To address this problem, researchers have proposed numbers of anti-spam approaches from different perspectives, including legal means, working out corresponding acts to regulate email sending [5,6]; email protocol methods, improving the control strategies of email protocols [7,8]; simple techniques, such as address protection [9], black/white list [10,11], keywords filtering [12] and so on; and intelligent detection, considering the spam filtering problem as a typical two-class classification problem, which could be solved by the supervised machine learning methods [12–14]. Among all these anti-spam approaches, intelligent detection is the most effective and widely used. On the one hand, intelligent detection is highly automated and do not need much human intervention; On the other hand, intelligent detection has the characteristics of high accuracy, robustness and strong noise tolerance, and it can adapt to the dynamic changes of the emails' content and users' interests.

There are three main related research fields for intelligent spam detection as well as other classification or pattern recognition problems, namely feature selection, feature construction and classifier design, corresponding to the three core steps of intelligent spam detection. The purpose of feature selection lies in reducing the number of features to be further processed and the affect from possible noisy features, so as to reduce the computational complexity and enhance the categorization accuracy respectively. Several feature selection metrics have been proposed and proved to be effective, such as Information Gain (IG) [15], Document Frequency (DF) [16], Term Frequency Variance (TFV) [17], Chi Square (χ^2) [16], Odds Ratio (OR) [18], Term Strength (TS) [16] and so on. Feature construction approaches transform the set of features available into a new set of features by finding relationships between existing features and constructing feature vectors to represent samples. Bag-of-Words (BoW), also known as Space Vector Model, is the most widely used feature construction approach in spam detection [19]. Other feature construction approaches for spam detection have also been studied, like Sparse Binary Polynomial Hashing (SBPH) [20], Orthogonal Sparse Bigrams (OSB) [21], immune concentration based approaches [22–27] and term space partition (TSP) based approach [28] etc. Supervised machine learning methods have been successfully and widely applied for classifier design in spam detection, and the prevalent ones are introduced in Section 2.

This paper applies Stacked Auto-Encoders (SAE), one of the main types of deep neural networks, to intelligent spam detection. And presents a comparative study of SAE with other prevalent supervised machine learning methods to verify the effectiveness of deep learning on spam detection. Experiments were conducted on five benchmark corpora PU1, PU2, PU3, PUA and Enron-Spam to

investigate the performance of SAE and other machine learning methods. Accuracy and F_1 measure are selected as the main criteria in analyzing and discussing the results.

The rest of this paper is organized as follows: Section 2 introduces the prevalent machine learning methods that are applied in spam detection. Stacked Auto-Encoders is presented in detail in section 3. Section 4 gives the experimental results and corresponding analysis. Finally, we conclude the paper in Section 5.

2 Prevalent Machine Learning Methods

2.1 Naive Bayes

The Bayes methods compute the probability $P(C = c_k|X = x)$ that the sample x belongs to each category c_k and obtain the final category of sample x according to the maximum value of the probability that has been achieved.

$$P(C = c_k|X = x) = \frac{P(X = x|C = c_k)P(C = c_k)}{P(X = x)} \tag{1}$$

According to the Bayes formula shown in Eq.1, the key part of the Bayes methods is computing the probability $P(X = x|C = c_k)$. Naive Bayes (NB) is the most widely used Bayes method, and it assumes that the sample x is composed of multiple features w_j which are mutually independent in the calculation process, thus $P(X = x|C = c_k)$ could be achieved by computing $P(W = w_j|C = c_k)$. Sahami et al. [29] introduced NB into spam detection, and now it has been widely used in commercial spam filtering system and open source software of spam detection based on its simplicity in implementation and high accuracy.

2.2 Support Vector Machine

The core idea of Support Vector Machine (SVM) is to find the optimal hyperplane and make the classification margin maximized. The targets of training process is maximizing the classification margin and minimizing the structural risk, and obtaining weight vector of the optimal hyperplane by calculation on the training set. For linearly inseparable issues, SVM makes it linearly separable by mapping the training data from the original space to a higher-dimensional space with kernel functions and computes corresponding optimal hyperplane. Drucker et al. [30] applied SVM to spam detection and achieved better performance compared with Ripper and Rocchio. In addition, best performance of SVM was achieved when boolean BoW is employed as feature construction approach other than multi-value BoW.

2.3 Decision Tree

Decision Tree (DT) constructs a tree from top to bottom according to the predefined sequence of attributes, where nodes corresponds to attributes and edges

corresponds to attribute values. Each path from the root to the leaves could be seen as a rule. Selecting the sequence of attributes based on IG is one of the commonly used methods in DT. The famous DT algorithms are ID3 and C4.5 etc. Carreras et al. [31] applied DT to spam filtering and adopted RLM distance other than IG for attributes selection. Currently, DT is often used as a weak learner of Boosting methods due to its mediocre performance.

2.4 Artificial Neural Network

Artificial Neural Network (ANN) is proposed by taking inspiration from mechanism of biological neural networks and consists of a large number of interconnected artificial neurons. There are three types of neurons: the input layer neurons, hidden layer neurons and output layer neurons. In the learning (training) process, the connection weights of ANN are dynamically adjusted in accordance with the input and output values of the training data to approximate the mapping function of the input and output values. In the classification process, the input data transfer in the network layer by layer beginning from the input layer. The activation value of each neuron is calculated according to the predefined activation function and effect of each neuron in the classification is determined by the connection weights. Performance of ANN is mainly influenced by three factors: input and the activation function, network structure and connection weights. Clark et al. [32] adopted ANN to classify emails with a fully connected neural network, and used back-propagation (BP) algorithm for training. Experimental results showed that ANN could achieve better performance than NB and k-Nearest Neighbor.

2.5 Boosting

Boosting could be seen as a voting technology based on existing learning methods, other than a particular learning method itself. AdaBoost (Adaptive Boosting) is a typical Boosting method. The core idea of this method is giving more attention to the samples those are difficult to be classified in the learning process [33]. During the training process, weights of samples are dynamically adjusted in accordance with their classification results by the constructed classifiers, and the samples those are difficult to be classified would be selected for learning with greater probability when the new classifiers are built. Finally, new samples are weighted classified according to the performance of each classifier. Carreras et al. [31] applied AdaBoost to spam detection by using DT as base classifier, and AdaBoost achieved better performance than DT and NB in the experiment.

2.6 Random Forest

Random Forest (RF) samples repeatedly from the original sample set by utilizing the re-sampling method bootsrap and constructs Decision Tree model on each bootsrap sampling. Then the DT models constructed are combined to give the

prediction by voting. Koprinska et al. [17] applied RF to email classification and compared its performance with that of other methods. Experimental results indicate that RF is promising approach for spam filtering and outperforms DT, SVM and NB, with DT and SVM being also more complex than RF.

3 Stacked Auto-Encoder

Artificial Neural Networks are traditional computational models for machine learning and pattern recognition. Former researches mainly focus on shallow neural networks (i.e. neural networks with one hidden layer or two hidden layers). Since deep neural networks have shown excellent performance in recent years, Deep Learning (DL) has become a hot topic in artificial intelligence. Deep Learning algorithms attempt to learn multiple levels of representation of increasing complexity or abstraction and deep multi-layer neural networks are the basic architectures of DL.

Stacked Auto-Encoder (SAE) is one of the main types of deep networks. It is a stacked ensemble of auto-encoders and has more excellent computational ability [34].

The structure of auto-encoder is show in Fig.1, which consists of three layers, namely input layer, code layer and reconstruction layer. The original input X enters at the input layer, and X is encoded to Y through forward-propagation in the neural network. Further, Y is decoded to X'. In auto-encoder, X' has the same dimensionality with X and is seen as a reconstruction of the original input X.

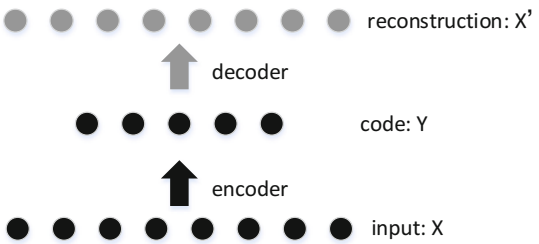

Fig. 1. Structure of Auto-encoder

Simply speaking, the auto-encoder transforms the input vector $X = (x_1, x_2, ..., x_n)$ to vector $Y = (y_1, y_2, ...y_m)$, where n indicates dimensionality of the input vector and m indicates dimensionality of the code vector. Next, the input vector X is reconstructed to X' from the code vector Y with the constraint that $|X - X'|$ is minimized. The training objective of this neural network is minimizing the reconstruction error, and the objective function is defined as follows:

$$J = \sum \|X - X'\| \tag{2}$$

where the sum operation is executed on all input samples. In this paper, the network is trained with the gradient descent BP algorithm, which is widely used in the training of artificial neural networks.

Actually, the code Y is a nonlinear abstract of original input data X and represents some features of X. SAE learns multiple levels of representation of the input vector X, in which the high layer encodes the low layer and each layer represents features of the input with increasing abstraction, and reconstructs X to X' from the last layer with constraint that $|X - X'|$ is minimized, as shown in Fig.2. SAE is a stacked ensemble of multiple auto-encoders, in which X is encoded to Y, Y is encoded to Z and Z is encoded to W successively. Reconstruction is taken in the reverse order. W is decoded to Z', Z' is decoded to Y' and Y' is decoded to X' successively. The training objective of SAE is the same with auto-encoder, which is minimizing the reconstruction error of X.

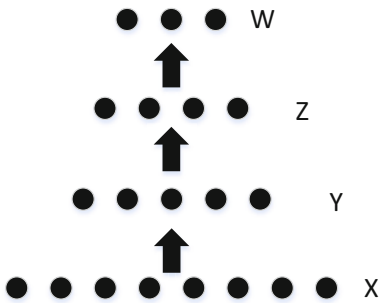

Fig. 2. Structure of Stacked Auto-Encoder

Back Propagation (BP) works well for networks with one or two hidden layers, while training deeper networks through BP yields poor results. SAE is trained by adopting the greedy layerwise pre-training [35,36], which greedily trains one layer at a time, exploiting an unsupervised learning algorithm for each layer. The auto-encoder X-Y is trained on the original data X first and transforms X to code Y. Then the auto-encoder Y-Z is trained the same as above based on Y and encodes Y to Z. Finally, the auto-encoder Z-W is trained on Z. After training of the three individual auto-encoders, the weights got is used to initialize the weight of SAE. The objective function is optimized by using BP algorithm.

Initialization strategy of BP weights is introduced in the training of SAE. The laywise pre-training of each auto-encoder could preliminarily determine the distribution of the initial data and make the wights of the neural network reflect the data characteristics. Initializing SAE with the weights got by pre-training could locate the initial solution close to the optimal solution and effectively reduce the convergence time. SAE possesses greater nonlinear capability than auto-encoder by laywise encoding of the original data. Each layer presents an

individual abstract of the original data and higher layer has higher level abstraction. SAE is considered having strong learning ability and able to mine the effective features of original data sufficiently. In this paper, we applied SAE in spam detection to verify the performance of DL in spam filtering.

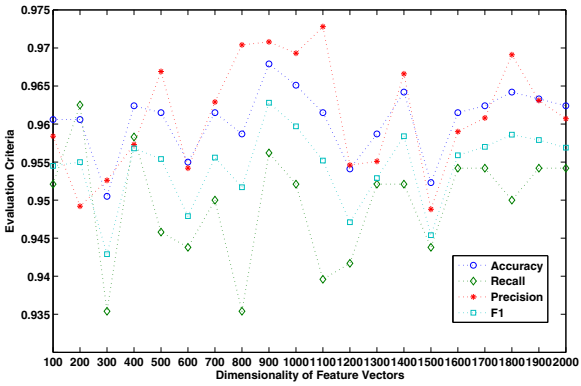

Fig. 3. Performance of AdaBoost with Varied Feature Vector Dimensionality on PU1

4 Experiments

4.1 Experimental Setup

In the experiments, Information Gain (IG) [15] and Bag-of-Words (BoW) [19] are selected as feature selection strategy and feature construction approach respectively for transforming email samples into feature vectors. IG is the most widely employed feature goodness criterion in machine learning area. It measures the number of bits of information obtained for class prediction by knowing the presence or absence of a certain feature in a sample. When applied in spam detection, IG of term t_i is calculated as

$$IG(t_i) = \sum_{c\in(s,h)} \sum_{t\in(t_i,\bar{t_i})} P(t,c) \log \frac{P(t,c)}{P(t)P(c)} \tag{3}$$

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. BoW, also known as Space Vector Model, is one of the most widely used feature construction approaches in spam detection. It transforms an email m to a n-dimensional feature vector $\boldsymbol{x} = [x_1, x_2, ..., x_n]$ 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. We take binary representation, where x_i is equal to 1 when t_i occurs in m, and 0 otherwise.

Experiments were conducted on PU1, PU2, PU3, PUA [37] and Enron-Spam [38], which are all benchmark corpora widely used for effectiveness evaluation in spam detection. Among them, 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; and 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.

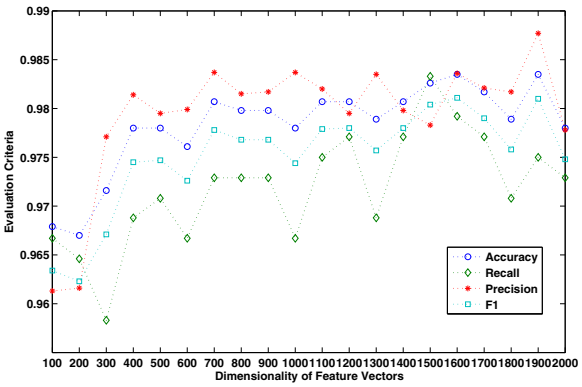

Fig. 4. Performance of Random Forest with Varied Feature Vector Dimensionality on PU1

In addition, SAE was implemented in MATLAB by utilizing the toolbox for deep learning [39], and a six-layer neural network is employed, in which the computational elements of each layer are 2000, 500, 250, 125, 10, 1 respectively. WEKA toolkit [40] was utilized in implementation of the machine learning models selected for comparison, namely Naive Bayes, Support Vector Machine, C4.5, Multilayer Perceptron (MLP), AdaBoost (C4.5 is selected as base learner) and Random Forest. Since determining the number of terms (features) selected for further classification, selection of dimensionality of feature vectors for each machine learning method can not only affect the computational complexity, but also influence the classification performance. Dimensionality of feature vectors for most of the techniques above are set in accordance with the previous researches, where that of MLP is set to 2000, the same as SAE [35] and those of NB, SVM and C4.5 are set to 500 [41]. While the dimensionality of feature vectors for AdaBoost and RF are investigated by conducting experiments on the relatively smaller corpus PU1, and set to 900 and 1600 respectively, as shown in Fig.3 and Fig.4. 10-fold cross validation was utilized on PU corpora and 6-fold

cross validation on Enron-Spam according to the number of parts each of the corpora has been already divided into. Accuracy and F_1 measure are the main evaluation criteria, as they can reflect the overall performance of spam detection.

4.2 Performance Comparison

Table 1 to 5 show the performance of different machine learning methods in spam detection when incorporated with IG and BoW. As can be seen, SAE performs the best in most of the cases in terms of both accuracy and F_1 measure, except that it has a similar performance with RF on PU1 (as mentioned above, we take accuracy and F_1 measure as comparison criteria without focusing on precision and recall, which are incorporated into the calculation of F_1 measure and can be reflected by F_1 measure). This indicates that SAE can work well and outperform the current machine learning methods in spam detection, verifying the effectiveness of deep learning in this area and endowing it with application in the real world.

By comparing the performance of the above machine learning methods between different corpus, we can see that SAE can achieve similar and relatively higher accuracy and F_1 measure on all of the corpus selected for the experiments (as well as MLP and SVM), demonstrating that SAE possesses good stability

Table 1. Performance comparison of SAE with other prevalent machine learning techniques on PU1

Method	Precision(%)	Recall(%)	Accuracy(%)	F_1(%)
NB	97.74	82.50	91.47	89.37
C4.5	91.24	89.17	91.28	90.01
SVM	96.19	95.62	96.33	95.77
AdaBoost	97.08	95.62	96.79	96.28
RF	98.36	97.92	**98.35**	**98.11**
MLP	97.99	97.50	97.98	97.68
SAE	98.16	97.71	98.17	97.89

Table 2. Performance comparison of SAE with other prevalent machine learning techniques on PU2

Method	Precision(%)	Recall(%)	Accuracy(%)	F_1(%)
NB	82.67	70.71	90.70	75.34
C4.5	79.26	71.43	89.72	73.96
SVM	90.99	78.57	94.08	83.92
AdaBoost	91.74	75.71	93.66	82.23
RF	97.46	65.00	92.68	76.83
MLP	90.85	89.29	95.91	89.57
SAE	93.29	88.57	**96.34**	**90.37**

Table 3. Performance comparison of SAE with other prevalent machine learning techniques on PU3

Method	Precision(%)	Recall(%)	Accuracy(%)	F_1(%)
NB	92.75	78.63	87.72	84.93
C4.5	91.10	91.76	92.25	91.34
SVM	95.44	93.96	95.33	94.67
AdaBoost	95.54	94.56	95.62	95.02
RF	97.50	95.66	96.97	96.55
MLP	96.72	95.66	96.63	96.17
SAE	96.77	97.14	**97.24**	**96.91**

Table 4. Performance comparison of SAE with other prevalent machine learning techniques on PUA

Method	Precision(%)	Recall(%)	Accuracy(%)	F_1(%)
NB	95.39	94.21	94.65	94.63
C4.5	87.02	92.63	88.68	89.30
SVM	91.84	94.39	92.63	92.87
AdaBoost	91.08	97.02	93.42	93.80
RF	92.15	97.02	94.04	94.36
MLP	94.24	97.02	95.35	95.49
SAE	94.49	97.90	**95.88**	**96.04**

Table 5. Performance comparison of SAE with other prevalent machine learning techniques on Enron-Spam

Method	Precision(%)	Recall(%)	Accuracy(%)	F_1(%)
NB	77.74	98.41	87.45	86.10
C4.5	82.88	97.07	90.33	89.02
SVM	89.64	98.74	94.63	93.86
AdaBoost	89.13	98.78	94.25	93.57
RF	91.46	99.28	96.06	95.11
MLP	92.70	98.71	96.23	95.54
SAE	94.90	98.95	**97.49**	**96.84**

and robustness due to its strong learning ability, which is concerned more in the real world application. While the others usually occur with the phenomenon that the performance declines significantly on some of the corpus selected.

In addition, NB is a simple and efficient method which can recognize majority of the samples, hence it is widely used in the real spam filtering systems. AdaBoost and RF are both ensemble methods based on weak learners and obtain better performance than NB and C4.5 (which are weak learners) in the experiments.

It is worth mentioning that the training of SAE is really time consuming as well as MLP (and AdaBoost on large corpus). However, this could be settled by the offline training of real world spam filters and strong computational ability of modern computers and servers.

5 Conclusion

In this paper, we applied SAE to spam detection and compared its performance with the prevalent machine learning techniques those are commonly used in this area. Comprehensive experiments were conducted on public benchmark corpus and a six-layer neural network was employed to investigate the performance of SAE. The results demonstrate that SAE not only outperforms other machine leaning methods in terms of classification accuracy and F_1 measure, but also possesses stronger stability and robustness. This verifies the effectiveness of deep learning in spam filtering and endows it with application in real world meanwhile. In future work, we intend to further apply the other types of deep networks in spam detection and construct novel feature construction models in accordance with the characteristics of spam problem and advantages of different deep networks.

Acknowledgments. This work was supported by the Natural Science Foundation of China (NSFC) under grant no. 61375119, 61170057 and 60875080, and partially supported by National Key Basic Research Development Plan (973 Plan) Project of China with grant no. 2015CB352300.

References

1. Cranor, L., LaMacchia, B.: Spam!. Communications of the ACM **41**(8), 74–83 (1998)
2. Research, F.: Spam, spammers, and spam control: A white paper by ferris research. Technical Report (2009)
3. Corporation, S.: Internet security threat report: 2014. Technical Report (2014)
4. Cyren: Internet threats trend report: October 2014. Technical Report (2014)
5. Lugaresi, N.: European union vs. spam: a legal response. In: Proceedings of the First Conference on E-mail and Anti-Spam (2004)
6. Moustakas, E., Ranganathan, C., Duquenoy, P.: Combating spam through legislation: a comparative analysis of us and european approaches. In: Proceedings of the Second Conference on Email and Anti-Spam, pp. 1–8 (2005)
7. Marsono, M.N.: Towards improving e-mail content classification for spam control: architecture, abstraction, and strategies. PhD thesis, University of Victoria (2007)
8. Duan, Z., Dong, Y., Gopalan, K.: Dmtp: Controlling spam through message delivery differentiation. Computer Networks **51**(10), 2616–2630 (2007)
9. Hershkop, S.: Behavior-based email analysis with application to spam detection. PhD thesis, Columbia University (2006)
10. Sanz, E., Gomez Hidalgo, J., Cortizo Perez, J.: Email spam filtering. Advances in Computers **74**, 45–114 (2008)

11. Heron, S.: Technologies for spam detection. Network Security **2009**(1), 11–15 (2009)
12. Cormack, G.: Email spam filtering: A systematic review. Foundations and Trends in Information Retrieval **1**(4), 335–455 (2007)
13. Carpinter, J., Hunt, R.: Tightening the net: A review of current and next generation spam filtering tools. Computers & security **25**(8), 566–578 (2006)
14. Kotsiantis, S.: Supervised machine learning: A review of classification techniques. Informatica **31**, 249–268 (2007)
15. Yang, Y.: Noise reduction in a statistical approach to text categorization. In: Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 256–263. ACM (1995)
16. Yang, Y., Pedersen, J.: A comparative study on feature selection in text categorization. In: Machine Learning-International Workshop Then Conference-, Morgan Kaufmann Publishers, INC, pp. 412–420 (1997)
17. Koprinska, I., Poon, J., Clark, J., Chan, J.: Learning to classify e-mail. Information Sciences **177**(10), 2167–2187 (2007)
18. Shaw, W.: Term-relevance computations and perfect retrieval performance. Information Processing & Management **31**(4), 491–498 (1995)
19. Guzella, T., Caminhas, W.: A review of machine learning approaches to spam filtering. Expert Systems with Applications **36**(7), 10206–10222 (2009)
20. Yerazunis, W.: Sparse binary polynomial hashing and the crm114 discriminator. the Web (2003). http://crm114.sourceforge.net/CRM114paper.html
21. Siefkes, C., Assis, F., Chhabra, S., Yerazunis, W.S.: Combining winnow and orthogonal sparse bigrams for incremental spam filtering. In: Boulicaut, J.-F., Esposito, F., Giannotti, F., Pedreschi, D. (eds.) PKDD 2004. LNCS (LNAI), vol. 3202, pp. 410–421. Springer, Heidelberg (2004)
22. Tan, Y., Deng, C., Ruan, G.: Concentration based feature construction approach for spam detection. In: Neural Networks, 2009. IJCNN 2009. International Joint Conference on, pp. 3088–3093. IEEE (2009)
23. Ruan, G., Tan, Y.: A three-layer back-propagation neural network for spam detection using artificial immune concentration. Soft Computing-A Fusion of Foundations, Methodologies and Applications **14**(2), 139–150 (2010)
24. Zhu, Y., Tan, Y.: Extracting discriminative information from e-mail for spam detection inspired by immune system. In: 2010 IEEE Congress on Evolutionary Computation (CEC), pp. 1–7. IEEE (2010)
25. Zhu, Y., Tan, Y.: A local-concentration-based feature extraction approach for spam filtering. IEEE Transactions on Information Forensics and Security **6**(2), 486–497 (2011)
26. Mi, G., Zhang, P., Tan, Y.: A multi-resolution-concentration based feature construction approach for spam filtering. In: The 2013 International Joint Conference on Neural Networks (IJCNN), pp. 1–8. IEEE (2013)
27. Gao, Y., Mi, G., Tan, Y.: An adaptive concentration selection model for spam detection. In: Tan, Y., Shi, Y., Coello, C.A.C. (eds.) ICSI 2014, Part I. LNCS, vol. 8794, pp. 223–233. Springer, Heidelberg (2014)
28. Mi, G., Zhang, P., Tan, Y.: Feature construction approach for email categorization based on term space partition. In: The 2013 International Joint Conference on Neural Networks (IJCNN), pp. 1–8. IEEE (2013)
29. Sahami, M., Dumais, S., Heckerman, D., Horvitz, E.: A bayesian approach to filtering junk e-mail. In: Learning for Text Categorization: Papers from the 1998 Workshop, Madison, Wisconsin, vol. 62, pp. 98–105. AAAI Technical Report WS-98-05 (1998)

30. Drucker, H., Wu, D., Vapnik, V.: Support vector machines for spam categorization. IEEE Transactions on Neural Networks **10**(5), 1048–1054 (1999)
31. Carreras, X., Marquez, L.: Boosting trees for anti-spam email filtering. arXiv preprint cs/0109015 (2001)
32. Clark, J., Koprinska, I., Poon, J.: Linger-a smart personal assistant for e-mail classification. In: Proc. of the 13th Intern. Conference on Artificial Neural Networks (ICANN 2003), Istanbul, Turkey, pp. 26–29, June 2003
33. Rokach, L.: Ensemble-based classifiers. Artificial Intelligence Review **33**(1), 1–39 (2010)
34. Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., Manzagol, P.A.: Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. The Journal of Machine Learning Research **11**, 3371–3408 (2010)
35. Hinton, G.E., Salakhutdinov, R.R.: Reducing the dimensionality of data with neural networks. Science **313**(5786), 504–507 (2006)
36. Bengio, Y., Lamblin, P., Popovici, D., Larochelle, H., et al.: Greedy layer-wise training of deep networks. Advances in neural information processing systems **19**, 153 (2007)
37. Androutsopoulos, I., Paliouras, G., Michelakis, E.: Learning to filter unsolicited commercial e-mail. "DEMOKRITOS", National Center for Scientific Research (2004)
38. Metsis, V., Androutsopoulos, I., Paliouras, G.: Spam filtering with naive bayes-which naive bayes. In: Third Conference on Email and Anti-Spam (CEAS), vol. 17, pp. 28–69 (2006)
39. Palm, R.B.: Prediction as a candidate for learning deep hierarchical models of data. Technical University of Denmark, Palm 25 (2012)
40. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.: The weka data mining software: an update. ACM SIGKDD Explorations Newsletter **11**(1), 10–18 (2009)
41. Zhu, Y., Mi, G., Tan, Y.: Query based hybrid learning models for adaptively adjusting locality. In: The 2012 International Joint Conference on Neural Networks (IJCNN), pp. 1–8. IEEE (2012)