SPAM FILTERING SECURITY EVALUATION FRAMEWORK USING SVM, LR AND MILR

Kunjali Pawar[1] and Madhuri Patil[2]

[1]M.E. Student, Dr. D.Y.Patil School of Engg. And Technology, Lohegaon, Pune, Savitribai Phule Pune University, India.
[2]Assistant Professor, Dr. D.Y.Patil School of Engg. And Technology, Lohegaon, Pune, Savitribai Phule Pune University, India.

ABSTRACT

The Pattern classification system classifies the pattern into feature space within a boundary. In case adversarial applications use, for example Spam Filtering, the Network Intrusion Detection System (NIDS), Biometric Authentication, the pattern classification systems are used. Spam filtering is an adversary application in which data can be employed by humans to attenuate perspective operations. To appraise the security issue related Spam Filtering voluminous machine learning systems. We presented a framework for the experimental evaluation of the classifier security in an adversarial environments, that combines and constructs on the arms race and security by design, Adversary modelling and Data distribution under attack. Furthermore, we presented a SVM, LR and MILR classifier for classification to categorize email as legitimate (ham) or spam emails on the basis of thee text samples.

KEYWORDS

Adversary Model, Multiple Instance Logistic Regression, Pattern Classification, Security Evaluation, Spam Filtering

1. INTRODUCTION

This Machine learning systems bid an unparalleled resilience in acting with emerging input in a variation of applications, such as Intrusion Detection Systems (IDS) [1] and the spam filtering of e-mails. Whenever machine learning is used to prevent illegal or unsanctioned activity [2] and there is an economic incentive, adversaries will attempt to avoid the stability provided. Constraints on how adversaries can employs the training data (TR) and test data (TS) for classifiers used to encounter incredulous behaviour make problems in this area tractable and interesting. Pattern classification has earned eminence in different fields which contains security concerned applications like the Spam Filtering, the Network Intrusion Detection System (NIDS), and Biometric Authentication to distinguish between the legitimate and malicious samples [3].

In specific, there are three main clear issues can be recognized: (a) examining the weaknesses (vulnerabilities) of classification algorithms, and the corresponding attacks; (b) creating the novel methodologies to assess the classifier security under these attacks, which is not possible using the classical performance evaluation methodologies; (c) establishing the design methods [4] to guarantee the classifier security in an adversarial environment. The goal of attacker is to defeat the normal process of spam filters so that they can send spams [5].

DOI:10.5121/ijcax.2016.3302

The respite of the paper is unionized as follows: The section II, scrutinize about the problem definition on the Security Evaluation. The section III, discuss a proposed system framework for Spam Filtering Security Evaluation. The section IV scrutinizes the algorithms and evaluation of these algorithms using examples. The remainder of section V covers expected results of classifiers. The section VI, summarize the conclusion and the future scope.

2. PROBLEM STATEMENT

2.1. Problem statement

- Existing methods address one of the main open issues of evaluating at design phase the security of pattern classifiers.
- Even though the design phase of secure classifiers is a different issue than security evaluation, existing framework could be exploited to this end. For instance spam filtering existing system considers SVM and LR classifier.
- To apply an empirical security evaluation framework and provide security to Spam Filtering application and use best classifier in our framework.

2.2. Solving approach

The proposed system focuses on multiple instance logistic regression (MILR) strategy. In the recommended strategy, emails are treated as bags of multiple instances [6] and a logistic model at the instance level is indirectly learned by exploiting the bag level binomial log-likelihood function [14].

3. PROPOSED FRAMEWORK

The contribution of this paper is-

- Classification of email using SVM, LR and MILR classifiers.
- Intent to increase classification results, classifier called Multiple Instance Logistic Regression (MILR) is used.
- MILR differs from a single instance supervised learning [7], such that by splitting an email into several instances, a MI learner will be capable to identify the spam part of the text mail even though text mail has been injected with good words which solve the efficiency issue for GWI attack [14].

The data distribution gives the training data and testing data separately [8]. The testing data can be manually generated by the user during the compose emails. The assumptions can be given with the help of Adversary Model [9]. Modelling the adversary is dependent on the attack scenarios. It consists of goal, knowledge, capability, strategies of the adversary as shown in figure 1.

The classification application can be authenticated by authenticator. It consists of three classifiers like SVM, LR and MILR. The classifiers acts like an algorithms. These classifiers give its classification results either the email is spam or legitimate or normal. The training data can be already trained by admin. This user can be an attacker or an authorized person. For this application user can perform classification techniques (SVM, LR, and MILR) by means of analysis. This training region subsists of all types of mails such as spam or ham. For classification of testing part/mail using training part, different classifiers are used. SVM, LR and MILR classifiers are used for classification and result analysis [11].

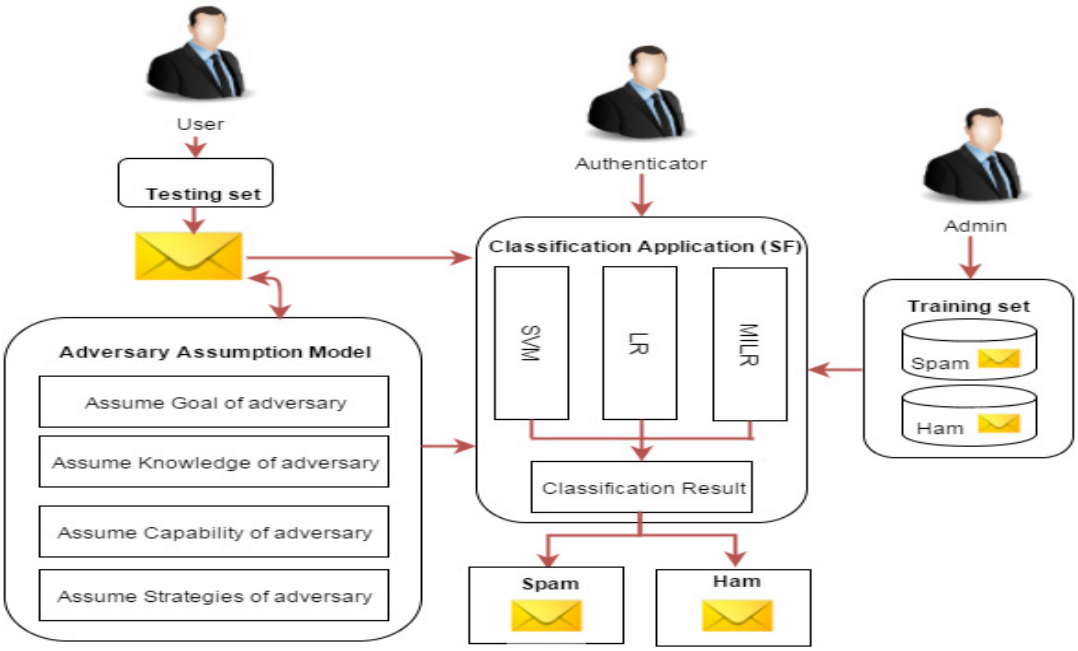


Figure1. Security Evaluation Framework using Spam Filtering

The Pattern classifier classifies or distinguishes the pattern (which is combination of feature which can be characterized by the individuals) into the feature space or word space within a boundary [10], [12], [13]. The goal is to partition the feature or word space in the class labeled decision regions. Therefore, for the ambition of model selection considers that the developer wants to appoint a Support Vector Machine (SVM) with a linear kernel, a Logistic Regression (LR) classifier, and Multiple Instance Logistic Regression (MILR). In the proposed system, for the classification purpose SVMs are actualized with the LibSVM, Logistic Regression (LR) and Multiple Instance Logistic Regression (MILR) is used for practical analysis.

4. ALGORITHMS WITH EXAMPLE

Firstly, a framework is presented for the empirical evaluation of classifier based on simulation of potential attack scenarios. The existing system considers SVM and LR classifiers [11]. The proposed system focuses on multiple instance logistic regression (MILR) strategy.

4.1. SVM Classifier

Algorithm: SVM classifier

Input:

Set of email data $D = \{d_1, d_2, .., d_n\}$,
(Combination of positively and negatively labeled data)

Output: Classified email data (Spam/Ham)

Process:

Step1: Distribution of positively and negatively labeled data according to features.

Step2: Compute Mapping function Ø().

Step3: Set bias as 1 to every vectors

Step4: Compute dot (.) products equations

Step5: Calculate the value of discriminate Hyperplanes α1, α2, α3,..

Step6: Predict positive and negative samples

Example:

Figure 2 shows the evaluation of SVM classifier with an example.

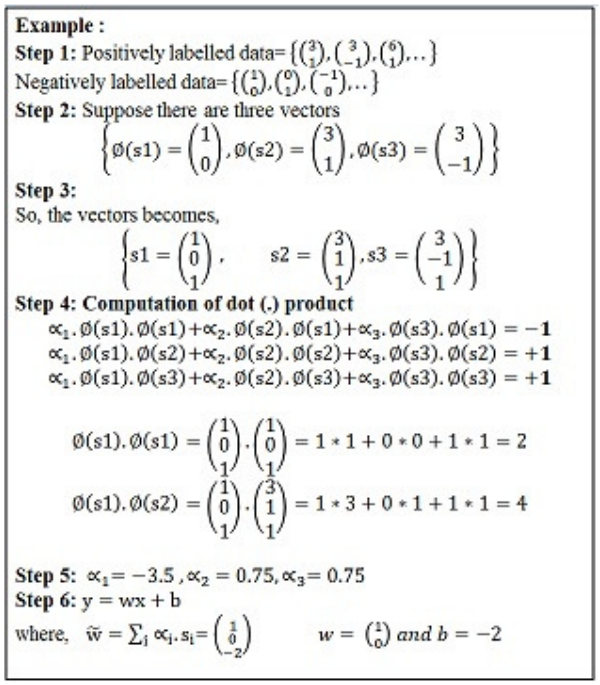


Figure2. SVM example

4.2. LR Classifier

Algorithm: LR classifier

Consider y= 1, if word spam and y=0, if word is not Spam (Ham).

For Spam Filtering, classification of words as spam or ham

The probability of word is

$$P(word) = (P(word|spam).P(spam) + P(word|ham).P(ham)$$

The probability for spam per word is

$$P(y = 1 \mid word) = \frac{P(word|spam).P(spam)}{P(word)}$$

The probability for ham per word is

$$P(y = 0 \mid word) = \frac{P(word|ham).P(ham)}{P(word)}$$

And

$$P(ham) = 1 - P(spam)$$

Example:

Figure 3 shows the evaluation of LR classifier with an example.

Example

Data- Hate, spam
Care, ham
Vulgar, spam

If word is Hate
The probability of Hate word is

$$P(Hate) = (P(hate|spam).P(spam) + P(Haate|ham).P(ham)$$
$$= 1 * \frac{1}{2} + 0 * \frac{1}{2} = \frac{1}{2}$$

The probability for spam per word (Hate) is

$$P(y = 1 \mid Hate) = \frac{P(Hate|spam).P(spam)}{P(Hate)}$$
$$= \frac{1 * \frac{1}{2}}{\frac{1}{2}} = 1$$

The probability for ham per word (Hate) is

$$P(y = 0 \mid Hate) = \frac{P(Hate|ham).P(ham)}{P(Hate)}$$
$$= \frac{0 * \frac{1}{2}}{\frac{1}{2}} = 0$$

Figure3. LR example

4.2. MILR Classifier

Algorithm: MILR classifier

Input:
Set of email data $D = \{d_1, d_2, .., d_n\}$,
$Pr(Y_i = 1|D_i)$ Be probability that the i[th] email is Positive,
$Pr(Y_i = 0|D_i)$ Be probability that the i[th] email is Negative

X_{ij} Is a vector of the word frequency counts (or *tf-idf* weight) of unique terms in every email.

Process:

Step 1: Binomial log-likelihood function is:

$$D_L = \sum_{i=1}^{m} [Y_i log Pr(Y_i = 1/D_i) + (1 - Y_i) log Pr(Y_i = 0/D_i)$$

Step 2: In a single instance setting, probability $Pr(Y_i = 1|X_i)$ has sigmoid response function as:

$$Pr(Y_i = 1|X_i) = \exp(p.X_i + b) / (1 + \exp(p.X_i + b))$$

Step 3: In multiple instances, setting estimate the instance-level class probabilities $Pr(Y_{ij} = 1|X_{ij})$ has a sigmoidal response function as:

$$Pr(Y_{ij} = 1|X_{ij}) = \exp(p.X_{ij} + b) / (1 + \exp(p.X_{ij} + b))$$

Where, X_{ij} is the j[th] instance in the i[th] data, and p and b are the parameters that need to be estimated.

Step 4: In a single instance setting, probability $Pr(Y_i = 0|X_i)$ has sigmoid response function as:

$$Pr(Y_i - 0|D_i) - \prod_{j=1}^{n} pr(Y_{ij} - 0|X_{ij}) - \exp(-\sum_{j=1}^{n} (\log(1 + \exp(p.X_{ij} + b))))$$

Step 5: In multiple instance setting estimate the instance-level class probabilities $Pr(Y_{ij} = 0|X_{ij})$ has a sigmoidal response function as:

$$Pr(Y_{ij} = 0|X_{ij}) = 1/(1 + \exp(p.X_{ij} + b))$$

5. EXPECTED RESULTS

The experimental results of the classification of legitimate (normal) and Spam e-mails are computed using the accuracy of the classifiers. The input of the proposed system is the number of testing samples. The computations of results are based on different mails. Some mails are trained before test and some new mails are also tested in proposed system. The analysis shows that MILR approach perform better as compare to SVM and LR algorithm by considering parameters like time, mean absolute error, etc. The accuracy can be calculated using following formula,

$$\text{Classification accuracy: } Ac = \frac{N_c}{N_t}$$

Where, N_c= Number of words which are correctly classified

N_t = Total number of words

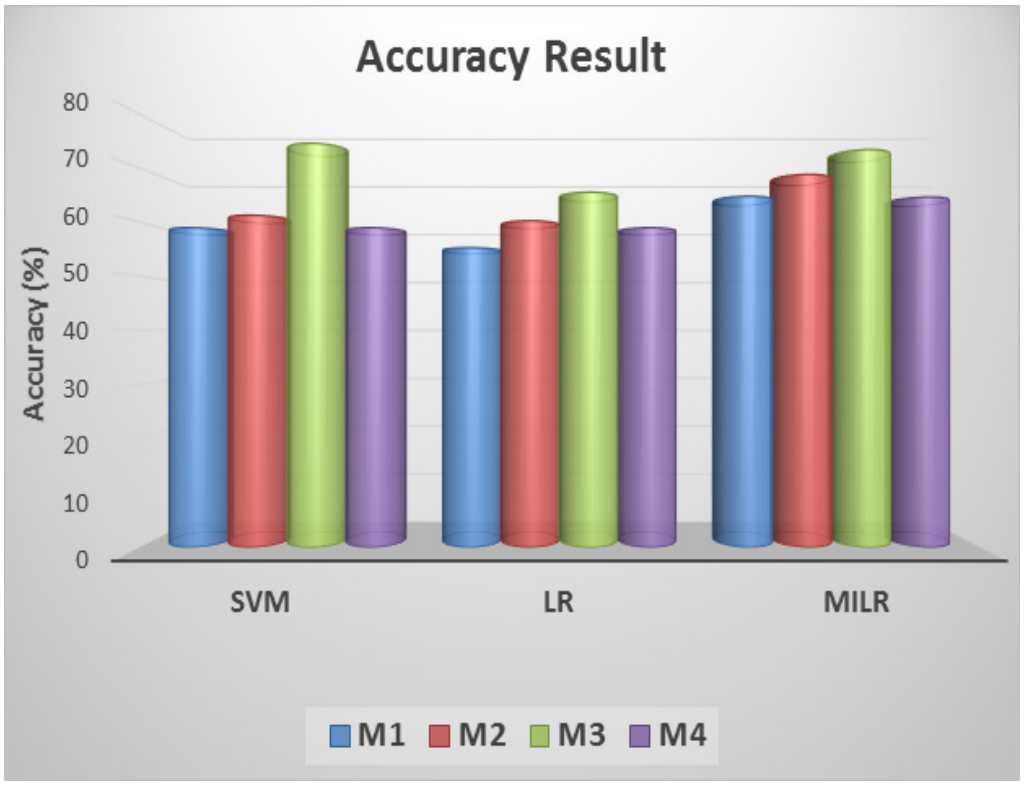
In this experiment, the efficiency of SVM, LR and MILR [11], [14] is evaluated to test the ability of classifiers [15]. Figure2 shows the output of accuracy on every classifier. Here we use the Spring Tool Suite (STS) with wampserver for database connectivity. Table 1 shows the accuracy result for queries M1, M2, M3 and M4 mail in terms of percentage (%).

M1: Code demonstrates how to get input from user.
M2: This is too dirty place. Please clean otherwise wash it.
M3: I hate you idiot. I will kill you idiot.
M4: The Indian Mujahideen claimed responsibility for the Jaipur bombings through an email sent to Indian media and declared open war against India.

TABLE I: Classification Accuracy

	M1	M2	M3	M4
SVM	58.93	61.24	74.49	58.93
LR	55.24	60.2	65.48	58.93
MILR	64.73	68.75	73.33	64.58

Figure2. Graphical representation of classification accuracy



6. CONCLUSION

This paper focused on experimental security evaluation of the pattern classifiers which improve prediction accuracy of spam filtering application. For classification and analysis three classifiers are used, are called SVM, LR and MILR. The proposed framework acquainted on a model of the adversary, and on a model of data distribution; accommodates an analytical approach for the training and testing sets generation that accredits security evaluation and can furnish the application distinct techniques. In the future, we will extend the data classification algorithm that will improve accuracy and performance of the system by means of spam detection.

ACKNOWLEDGEMENTS

We would like to thank the researchers as well as publishers for making their resources available and teachers for their guidance. We are thankful to the authorities of Savitribai Phule Pune University and concern members for their constant guidelines and support. We are also thankful to reviewer for their valuable suggestions and also thank the college authorities for providing the required infrastructure and support.

REFERENCES

[1] A. A. Cardenas, J.S. Baras, and K. Seamon, A Framework for the Evaluation of Intrusion Detection Systems, Proc. IEEE Symp. Security and Privacy, pp. 63-77, 2006.

[2] D.B. Skillicorn, Adversarial Knowledge Discovery, IEEE Intelligent Systems, vol. 24, no. 6, Nov./Dec. 2009.

[3] M. Barreno, B. Nelson, R. Sears, A.D. Joseph, and J.D. Tygar, Can Machine Learning be Secure? Proc. ACM Symp. Information, Computer and Comm. Security (ASIACCS), pp. 16-25, 2006.

[4] Kunjali Pawar and Madhuri Patil, A Review on Security Evaluation for Pattern Classifier against Attack, International Journal of Computer Applications (IJCA) Proceedings on National Conference on Advances in Computing NCAC-2015(4): 19-22, December 2015. (ISSN: 0975-8887).

[5] Y. Song, Z. Zhuang, W. C. Lee, H. Li, C.L. Giles and J. Li Q. Zhao, Real-Time Automatic Tag Recommendation, Proc. 31st Ann. Intl ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR 08), pp. 515-522, 2008.

[6] A. Kolcz and C.H. Teo, Feature Weighting for Improved Classifier Robustness, Proc. Sixth Conf. Email and Anti- Spam, 2009.

[7] P. Laskov and R. Lippmann, Machine Learning in Adversarial Environments, Machine Learning, vol. 81, pp. 115- 119, 2010.

[8] M. Barreno, B. Nelson, A. Joseph, and J. Tygar, The Security of Machine Learning, Machine Learning, vol. 81, pp. 121- 148, 2010.

[9] D.B. Skillicorn, Adversarial Knowledge Discovery, IEEE Intelligent Systems, vol. 24, no. 6, Nov./Dec. 2009.

[10] P. Laskov and M. Kloft, A Framework for Quantitative Security Analysis of Machine Learning, Proc. Second ACM Workshop Security and Artificial Intelligence, pp. 1-4, 2009.

[11] B. Biggio, G. Fumera, and F. Roli, Security Evaluation of Pattern Classifiers under Attack, IEEE Transactions On knowledge and Data engg., vol. 26, No. 4, April 2014.

[12] D. Lowd and C. Meek, Good Word Attacks on Statistical Spam Filters, Proc. Second Conf. Email and Anti-Spam, 2005.

[13] R.O. Duda, P.E. Hart, and D.G. Stork, Pattern Classification, Wiley-Interscience Publication, 2000.

[14] Z. Jorgensen, Y. Zhou, and M. Inge, A Multiple Instance Learning Strategy for Combating Good Word Attacks on Spam Filters, J. Machine Learning Research, vol. 9, pp. 1115-1146, 2008.

[15] Kunjali Pawar and Madhuri Patil, Pattern Classification under Attack on Spam Filtering, IEEE International Conference on Research in Computational Intelligence and Communication Networks (ICRCICN-2015), November 2015.

AUTHORS

Ms. Kunjali Pawar received Bachelor of Engineering degree in Computer Science & Engineering in 2014 and now pursuing Post Graduation (M.E.) in the department of Computer Engineering from Dr. D.Y.Patil School of Engineering and Technology in the current academic year 2015-16. She is now studying for the domain Data Mining and Information Retrieval as research purpose on Security Evaluation of Pattern Classifier against attack using Spam filtering during her academic year.

Prof. Mrs. Madhuri Patil. She is presently working as an Assistant Professor in the department of computer engineering, Dr. D. Y. Patil School of Engineering and Technology, Pune, Maharashtra, India. She has 6 years experience in teaching field and her research area is Data Mining and Information Retrieval.