Contemporary Engineering Sciences, Vol. 8, 2015, no. 12, 533 - 538
HIKARI Ltd, www.m-hikari.com
http://dx.doi.org/10.12988/ces.2015.5269

An Enhanced Approach for Secure Pattern Classification in Adversarial Environment

P. Udhaya

School of Computing, SASTRA University, Thanjavur, India

G. Manikandan

School of Computing, SASTRA University, Thanjavur, India

Abstract

The main objective of pattern classification in data mining is to relegate the similarities among data. When a pattern classifier is employed in adversarial environments like spam filtering, biometric authentication and network intrusion detection, the performance and utility of the application will diminish when it frails under attack. In this work, for effective classification an enhanced method is proposed to ensure the officiation of classifier against attacks. To manifest this model, they are accustomed in adversarial application and performances of two classifiers namely SVM and LR are measured under simulated attack. From our experiments it is evident that LR classier has high accuracy in predicting the attacks than SVM.

Keywords: Data Mining, Classification, SVM, LR, Spam Filtering

1 Introduction

In data mining, classification is one of the methods for performing data analysis that identifies the substantial data classes which can be used in predicting the future data trends [6]. Pattern classification has gained grandness in many fields including security concerned applications such as spam filters, network intrusion detection, and biometric authentication to secernate the legitimatized and malicious sample. The input data might be evaded by the adversaries to diminish the officiation of the classifier. This endures a slips tram between the classifier and the attack contriver. Examples of the attack eccentrics are spoofing attack which fares its frustration on the classifier by succumbing the false data to the biometric system, interpolates the email with the spam words and castrates the packets or files in networks to campaign attacks.

With respect to pattern classification, three issues have to be addressed. First issue focuses on various types of attacks that occur in the process of classification. Second issue is to acquire a method for making the classifier secure. Finally a method needs to be developed to ensure the working of classifier in adversarial environment.

Most of the works in the existing literature focus on identifying only integrity violation. In this work, the focus is on to train the classifier to identify availability violation in addition to integrity violation.

The remainder of this paper is organised as follows: Section-2 comprises the literature survey and the proposed system is given in section-3. Section-4 represents the architecture of the model along with the details of the data set. Experimental results are tabulated in Section-5 and a brief conclusion is given in Section-6.

2 Literature Survey

In [1], a new mechanism for enhancing the security of machine learning algorithms is suggested by analysing the taxonomy of violation goals which are based on the error rates. The availability violation makes attacks on legitimate samples and integrity violation causes malicious to be modified as legitimate.

Taxonomy of different attacks and violations that are related to the spam filtering applications are discussed [2]. Causative attacks generally occur in trained dataset and exploratory attacks influence test dataset.

An extension of classical method is proposed in [3] which involve construction of framework that simulates a model of adversaries which empirically evaluate classifier robustness under various attacks.

The usage of LibSVM for Classification is proposed in [4]. LibSVM is a library package which utilizes a two-step process for classification. In the first step a dataset is used for training and to measure various parameters, which are followed by a testing mechanism to predict the number of test samples that matches with trained dataset.

Classical method for pattern classification involves training, testing, parameter extraction and model acquisition. Same distribution is used in constructing training and testing dataset [5].

A method was proposed to measure robustness of the classifier against attacks. Each data sample is assigned with feature weight as binary values and distributed data is evaluated based on variations in the feature weights [7].

In [8], what if-analysis is used as a simulation method for predicting the behaviour of the system against modelled hypotheses such as attack scenarios which measure the impact of independent variable on the dependent variable.

3 Proposed System

The principal theme of the proposed system is to develop an enhanced model which prevents adversarial attacks by utilizing a suitable data distribution. The data in the data set is labelled as train and test samples. These samples are to be classified using the classifier algorithms namely SVM and LR. The performances of these classifiers are measured under various simulated attacks. The proposed model considers both the modification of malicious samples and legitimized sample to predict false acceptance rate of classifiers. The resultant attack samples can be included into training dataset to improve the discriminatory of the classifier.

The dataset D_s is a collection of legitimized and malicious samples. Features of samples are extracted using classical resampling techniques such that Ds= [Fa, Ca]$_{a=1 \text{ to } n}$ where F_a represents the features of the samples and C_a represents the class label which describes to which class a sample belongs(Legitimized L or Malicious M). D (F, C) are acquired through the distribution of samples (F, C) and the general model is D (F, C) = D(C). D (F|C).

The training and testing dataset will be in the form of D_{TR} (F, C) = D_{TS} (F, C) = D (F, C) in the absence of any attacks. The occurrence of an attack can be expressed using a Boolean variable named X. This variable is expressed as (X=T) in the presence of attacks and as not(X=F) otherwise. The samples which have not been affected by the attack, use the variable X in the form D (F|C, X=F) = D (F, C). The steps in the proposed algorithm are summarized as follows:

Algorithm for Construction of Modified Dataset

```
Input      : Dataset $D_S$
               Number of samples S
               Distribution D (F|C). D (F|C, X)
Output     : Modified Dataset $D_M$
    1. $D_M \leftarrow \emptyset$
    2. For N(max),
    3. Sample class priors or Variable X or feature vector of samples
       D(F,C=M,X=T)
    4. Draw a sample $D_N$(F,C,X)
    5. $D_M \leftarrow D_M$ U $D_N$(F,C,X)
    6. Repeat Steps 2-5 for samples D(F,C=L,X=T)
End
Return $D_M$
```

The modified test dataset is generated by simulating the attacks on the dataset samples. For N (max) maximum values, the spam features are modified as legitimate and legitimate samples are budged as spam by changing the variable X or by changing feature set. This process is repeated by varying the value of N to generate datasets of different dimensions.

4 Proposed Architecture

The architecture of the proposed model is represented in Figure 1. In this approach the data in the dataset is partitioned into two categories namely: Training set and Test set. The contents of the test set is modified and given as an input to the classifier. The extracted features from the training set are used for the classifier training purpose. The outcome of this phase is given as an input to the classifier which evaluates this input with the modified testing set.

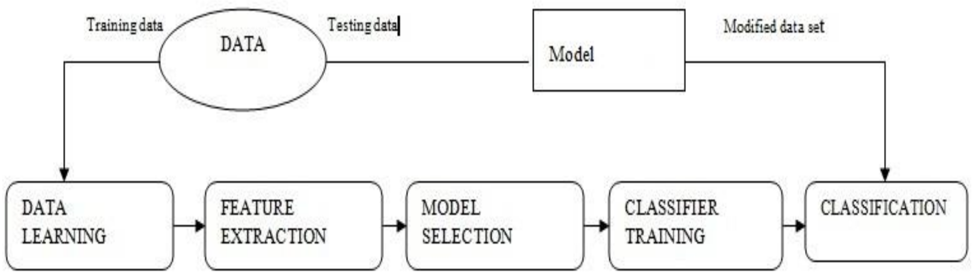

Fig 1: Proposed Model

5 Experimental Results

Experiments are carried out using Java 1.6 programming language and the outcomes are summarized in table 1. Dataset used for classification is TREC CORPUS 2007 email features. From our experimental outcomes it is evident that the performance of the LR classifier is higher than the SVM classifier. From our study it can be concluded that LR classifier is more accurate in the prediction of attacks than the SVM classifier in an adversarial environment.

Parameters	Accuracy of SVM Classifier	Accuracy of LR Classifier
Dataset with normal classification	75.80 %	74.50 %
Dataset without attack sample	84.29 %	84.76 %
Modified Dataset with attack sample	79.7788%	81.0442 %

Table-1 Performance of SVM and LR Classifiers

6 Conclusion

In this paper we have presented an enhanced model to improve the performance of the classification algorithms. From our experimental results it is observed the LR classifier has higher accuracy than SVM classifier in spam filtering. In future this research can be extended by utilizing the data model in other domains such as biometric authentication system, Intrusion Detection etc.

References

[1] M. Barreno, B. Nelson, A. Joseph, and J. Tygar. The Security of Machine Learning, Machine Learning, 81(2010), 121-148. http://dx.doi.org/10.1007/s10994-010-5188-5

[2] 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), 2006, 16-25. http://dx.doi.org/10.1145/1128817.1128824

[3] Battista Biggio, Giorgio Fumera, and Fabio Roli, Security Evaluation of Pattern Classifiers under Attack, IEEE Transactions on Knowledge and Data Engineering, 26(2014), 4, 984-996. http://dx.doi.org/10.1109/tkde.2013.57

[4] C. C. Chang, C. J. Lin, LibSVM: A Library for Support Vector Machines, 2001, http://www.csie.ntu.edu.tw/~cjlin/libsvm/

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

[6] Jiawei Han and Micheline Kamber, Data Mining: Concept and Techniques, Morgan Kauffman Publishers, 2010.

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

[8] S. Rizzi, What-If Analysis, Encyclopedia of Database Systems, Springer, 2009, 3525-3529. http://dx.doi.org/10.1007/978-0-387-39940-9_466

Received: March 11, 2015; Published: May 21, 2015