ISSN: 2321-7782 (Online)
Volume 3, Issue 3, March 2015
International Journal of Advance Research in Computer Science and Management Studies
Research Article / Survey Paper / Case Study
Available online at: www.ijarcsms.com

Proactive Prediction of features using Optimized SVM Classification

S. Surendar[1]
PG-Scholar / Dept. of C.S.E
SRM University, Chennai - India

R. Veeramani[2]
Professor / Dept. of I.T
SRM University, Chennai - India

Abstract: ***Many adversarial applications use Pattern classification systems in which data is manipulated by humans manually to know the operations underlying in it. These patterns are vulnerable to attacks which limits the performance and utility of the same. Using the optimized SVM instead of Patterns classification will be more secure and also improves the performance which outperforms the results from other classical methods of classification. This paper proposes an Optimized SVM based approach to evaluate the security against the email attacks. Also the reactive and proactive arms race between the adversary and classifiers has been discussed. The classification is done based on the rules or patterns using Dynamic Particle Swarm Optimization algorithm which can handle large amount of datasets. The PSO based approach has high rate of accuracy and performance compared with the other classification methods. We also compare the results of normal SVM and optimized SVM using PSO which shows better results.***

Keywords: Pattern classifiers, Rule based classification, Particle Swarm Optimization, SVM

I. INTRODUCTION

The classification systems based on the patterns use machine learning algorithms which are mostly used in the security related applications like authentications systems, intrusion detection systems etc. to find out the legitimate list of users and the content. These applications have intrinsic adversarial nature in which an intruder can manipulate the input data purposely to know the operation of the classifier. Well known examples are spoofing attack in biometric authentication systems; modifying network packets to evade intrusion detection systems; manipulating the content of spam emails to get them pass the spam filters.

It is said that, pattern classification system which follows old classical theory are more vulnerable to attacks since it does not take any adversarial settings into account. This issue should be addressed in a systematic way by using the rule based classification instead of pattern based classifiers. So the vulnerabilities of the classical algorithms needs to be analyzed and new method to assess the security of the classifiers needs to be developed to guarantee the security of the classifiers in the adversarial environments.

First thing is to improve the security in arms race between the classifier and the adversary. Instead of reacting to the observed attacks, the system should proactively anticipate the adversary by predicting the most appropriate and potential attacks through What if analysis. Through this method the countermeasures can be developed before the actual attack occurs which is mentioned as Principle of security by design. Then the practical guidelines should be defined to know the realistic attacks. Then training and test data also should be taken care separately.

II. CLASSIFICATION OF ATTACKS AGAINST PATTERN CLASSIFIERS

The taxonomy of attacks against the pattern classifiers are based on the two features [1]. First is the influence of attacks on the classifier and the second is the kind of the security violation it cause. The influence can be either causative or exploratory. If the influence undermines the learning algorithm to cause subsequent classifications, then it is said to be causative. If it exploits the knowledge of the well trained classifier to cause misclassifications without any effects on the learning algorithm then it is

exploratory. Both training and testing data will be affected by causative attacks but only testing data will be affected by exploratory attacks. The Security violation can be an integrity violation; availability violation or priority violation based on its characteristics. The integrity violation misclassifies the malicious samples as legitimate, whereas availability violation misclassifies the legitimate samples as malicious one. On the other side, the privacy violation allows the adversary to get all the confidential information from the classifier.

III. Feature Discovery and Reduction with Particle Swarm Optimization

The Particle Swarm Optimization (PSO) [5] is a rule discovery algorithm which uses a real-encoding way to discover the rules. It can be applied to both categorical and continuous attributes. It is a population based evolutionary technique which is inspired by the behavior simulation. PSO has been recognized widely and used in the non-linear functional optimization, pattern recognition, neural networks etc. In PSO algorithm, a swarm is referred to as a group of particles moving around in D-dimensional search space. The position of the i-th particle at the t-th iteration is being represented by $X_i^{(t)}=(x_{i1}, x_{i2}, \ldots, x_{iD})$ which is used to evaluate the quality of the particle. During the process of searching, the particle successively adjusts its position toward the global optimum based on the two factors: the best position encountered by itself (pbest) denoted as $Pi = (p_{i1}, p_{i2}, \ldots, p_{iD})$ and the best position encountered by the whole swarm (gbest) denoted as $P_g = (p_{g1}, p_{g2}, \ldots, p_{gD})$. Its velocity at the t-th iteration is given by $V_i^{(t)} = (v_{i1}, v_{i2}, \ldots, v_{iD})$. The following equations are used to calculate the position at next iteration

$$V_i^{(t)} = \lambda(\omega * V_i^{(t-1)} + c_1 * \text{rand}() * (P_i - X_i^{(t-1)}) + c_2 * \text{rand}() * (P_g - X_i^{(t-1)})) \quad (1)$$

$$X_i^{(t)} = X_i^{(t-1)} + V_i^{(t)} \quad (2)$$

where c_1 and c_2 are two positive constants defined as cognitive learning rate and social learning rate respectively; rand() is a random function in the range [0, 1] ;w is inertiafactor; and λ is constrictionfactor. In addition to this, the particle velocities are confined within [Vmin,Vmax]D. If velocity of the elements exceeds the threshold Vmin or Vmax, it is set equal to the respective threshold.

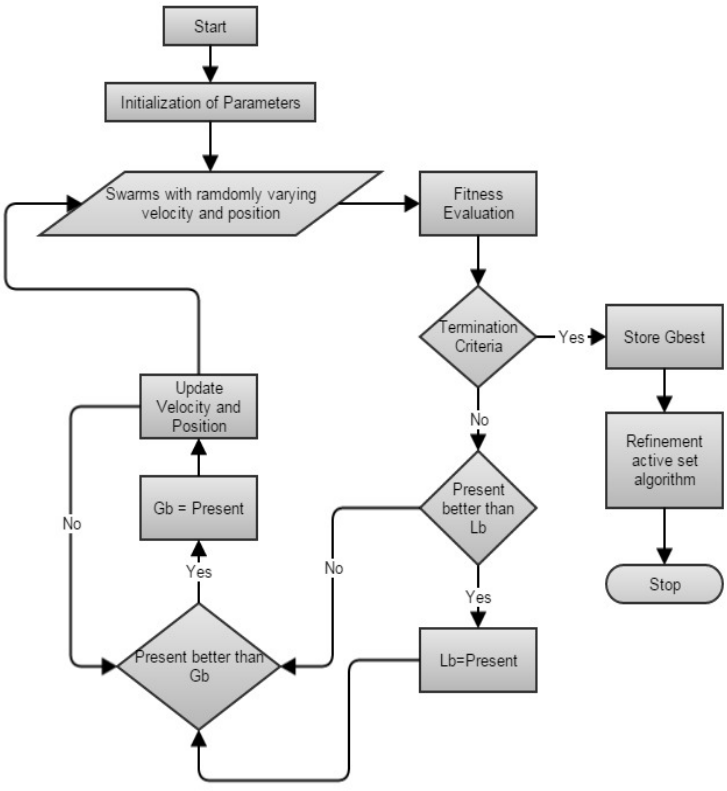

Fig.1 Functioning of PSO

The Pseudo code of the PSO algorithm is given below.

Input:

m: the swarm size; c1, c2 : positive acceleration constants;

w: inertia weight

MaxV: maximum velocity of particles

MaxGen: maximum generation

MaxFit: maximum fitness value

Output:

P_{gbest}: Global best position

Begin

Swarms $\{x_{id}, v_{id}\}$ =Generate (m); /*Initialize a population of particles with random positions and velocities on S dimensions */

Pbest(i)=0; i = 1,….,m,d=1,…..,S

Gbest = 0; Iter = 0;

While (Iter<MaxGen and Gbest < MaxFit)

{

For (every particle i)

{

Fitness(i)=Evaluate(i);

IF(Fitness(i)>Pbest(i))

{Pbest(i)=Fitness(i);$p_{id}=x_{id}$; d=1,…..,S}

IF(Fitness(i)>Gbest(i))

{Gbest(i)=Fitness(i);gbest=i; }

}

For (every particle i)

{

For(every d)

{

$v_{id} = w*v_{id} + c1*rand()*(p_{id}-x_{id})+c2*Rand()*(p_{gd}-x_{id})$

IF ($v_{id} > MaxV$) { $vi_d = MaxV$;}

IF ($v_{id} <- MaxV$) { $v_{id} =- MaxV$;}

$x_{id} = x_{id}+ v_{id}$

}

}

Iter=Iter+1;

}/* rand() and Rand() are two random functions in the range [0,1]*/

Return P_{gbest}

End

The PSO algorithm involves of just three steps, which are being replicated until stopping condition, they are as follows.

- Evaluate the fitness of each particle.
- Update individual and global best functions.
- Update velocity and position of each particle.

The PSO has certain advantages as easy to implement and computationally efficient algorithm. So it is applied in the selection of best feature training set using the cross validation and obtain the result using the SVM to predict the test set.

IV. IMPLEMENTATION OF PSO-SVM FOR SPAM EMAIL DATASET

Support vector machines (SVM) was used to detect spam by filtering the emails. SVMs are the learning system based statistical learning theory [3]. SVMs calculate a separate hyperplanes which maximizes the margin between data classes to produce good generalization results. SVMs were proved to be very efficient learning machine from various successful applications. Though SVMs has high performance, it have some limitations. Support Vector Machine has shown its power in the binary classification. It has strong theoretical foundation and learning algorithm. It shows better results in static data classification but the drawback is that it is memory and time consuming, that is expensive in computation, so it runs slow when the size of data is enormous. Also SVMs are more effective than other classical nonparametric classifiers like the nearest neighbour, neural networks and k-NN classifier with respect to computational time and stability to parameter setting and classification accuracy. But it is very weak to classify highly dimensional datasets with large number of features. In order to overcome this, an optimizer is needed to reduce the count of features before the subsequent feature subsets are available to SVM for better classification within stipulated time. There is a need improve the efficiency of the classification of SVM in terms of accuracy rate and time using Particle Swarm Optimization (PSO) by optimizing its feature selection process for the classification of email into spam emails or non-spam emails.

There are various libraries available for SVM implementation. To determine the Spam e-mail Dataset R software is used [4]. Kern lab is an integrated software package for support vector classification. Kern lab package[4] aims to provide the basic kernel functionality to the R user along with some other utility functions that are commonly used in kernel-based methods like a modern kernel-based algorithms and quadratic programming solver based on the functionality that the package provides. ksvm() in kernlab are a flexible SVM implementation which includes the most SVM formulations and kernels and allows for user defined kernels. It provides many useful options and features like a method for plotting, class probabilities output, cross validation error estimation. 10-fold cross validation is applied to validate and evaluates the provided solutions.

A boosting algorithm is used to train a set of classifiers on the data sets with completely different distributions, and joins them in an elementary way to reach a near-optimal performance. The boosting algorithms have been originally proposed by Schapire [8]. Schapire had proved that it is theoretically possible to convert a weak learning feature that performs slightly better than other random guessing into one achieves arbitrary accuracy. Boosting is a common method that is used to improvise the performance of any learning algorithms. Ensemble learning like bagging combines models built on re-samplings of various data yields a superior model. Instead of involving several distinct classifiers to boost the performance of classifier, we use a single classifier, PSO-SVM to achieve a near optimal performance.

V. EXPERIMENTAL WORK

The experiment conducted with the SVM classifier using MATLAB 7.7 [7].The experimental results obtained with SVM and optimized SVM technique are discussed below;

Results of Experiment with 1000 Email Messages

In the first phase, 1000 email messages are generated randomly. Among that 750 emails were used for training the classifier and 250 used to test it. The SVM classifies the 1000 email messages into spam and non-spam by finding a separating hyper-plane with the highest possible margin.

The accuracy (in %) and time obtained by SVM for the classification of the 1000 e-mail datasets is 68.34% while the time used is 6.33 seconds. Then the optimized SVM technique is used for the classification of the 1000 email messages into spam and non-spam by finding a separating hyper-plane with the highest possible margin. The accuracy (in %) and time used by optimized SVM technique for the classification of the 1000 e-mail datasets is 80.44% while the time used is 2.06 seconds.

Result of Experiment with 3000 Email Messages

In the second phase, 3000 e-mail message datasets were used. The training set has 2250 emails while the test set has 750 emails. Then the SVM classifies the 3000 email datasets into spam and non-spam. The accuracy (in %) and time used by SVM for the classification of the 3000 e-mail datasets is 46.71% while the time used is 60.16 seconds. Then the optimized SVM classification is done against the 3000 email datasets to split into spam and non-spam. The accuracy (in %) and time used by optimized SVM technique for the classification of the 3000 e-mail datasets is 90.56% while the time used is 0.56 seconds.

Results of Experiment with 6000 Email Messages

In the third phase, 6000 email datasets which is double than previous phase were used for the classification. The training set produced has 4500 email messages while the test set has 1500 email messages. The SVM classifies the 6000 email messages into spam and non-spam by finding a separating hyper-plane with the highest possible margin. The accuracy (in %) and time used by SVM for the classification of the 6000 email datasets is 18.02% while the time used is 91.47 seconds.

When the optimized SVM classifier is used for the classification of 6000 email messages into spam and non-spam, the accuracy is much better than ordinary SVM. The accuracy (in %) and time used by optimized SVM technique for the classification of the 6000 email datasets is 93.19% while the time used is 0.19 seconds.

Comparative Evaluation of SVM and Optimized SVM Techniques

After due evaluation of SVM and the optimized SVM technique, the result obtained is summarized and presented in Table 1.

TABLE 1

Experiment Results of SVM and PSO-SVM

Dataset	SVM		PSO-SVM	
	Classification Accuracy (%)	Computational Time (secs)	Classification Accuracy (%)	Computational Time (secs)
1000	68.34	6.33	80.44	2.06
3000	46.71	60.16	90.56	0.56
6000	18.02	91.47	93.19	0.19

Based on the result obtained from the experiment is summarized in Table 1, it can be observed that when dataset size is 1000, optimized SVM technique yields a classification accuracy of 80.44% in 2.06 seconds while SVM yields an accuracy of 68.34% in 6.33 seconds. This shows that the optimized SVM technique generated 12.1% higher accuracy in 4.27 seconds lesser computational time than SVM. When the dataset size is 3000, the result yielded for optimized SVM technique and SVM are 90.56% in 0.56 seconds and 46.71% in 60.16 seconds respectively. This shows that the optimized SVM technique gave 43.85% higher classification accuracy in the time frame of 59.6 seconds lesser computational time than ordinary SVM with an increase of 2000 emails. Finally, the classification accuracy and computational time for optimized SVM technique and SVM are 93.19% in 0.19 seconds and 18.01% in 91.47 seconds respectively. This shows that the optimized SVM technique produced 75.18% higher classification accuracy 91.28 seconds lesser computational time than SVM with an increase of 6000 emails. Therefore, the experimental results obtained shows that for every increase in dataset size, there is a reduction in the classification accuracy of SVM with a remarkable increase in computational time while the optimized SVM technique shows vice versa, thus it confirms the argument that SVM becomes weak and time consuming with large dataset size. Also PSO's ability to find an optimal set of feature weights that improve classification rate and accuracy.

Implication of the Results Obtained

SVM has been observed to consume a lot of computational resources and result in inaccurate classification in the phase of a large e-mail dataset. The result mentioned above indicates that optimized SVM technique has a remarkable improvement on computational time and classification accuracy over SVM in the face of a large e-mail dataset. This shows that the optimized SVM technique has overcome the drawbacks of SVM.

VI. CONCLUSION

This study presents a particle swarm optimization-based feature selection technique which is capable of searching for the optimal parameter values for SVM to obtain a subset of beneficial features that is very much required to detect the spam in the email datasets. PSO is applied to classify feature parameters for SVM classifier and optimize the feature subset selection in great manner. It eliminates the redundant and irrelevant features in the dataset of the spam messages which in turn reduces the feature vector dimensionality drastically. This helps SVM to select optimal feature subset from the resulting feature subset. This optimal subset of features is then adopted in both training and testing to obtain the optimal outcomes in classification. Comparison of the results with SVM classifier shows that the optimized SVM technique has better classification accuracy than SVM. The optimized SVM technique has shown a vast improvement over SVM in means of classification accuracy as well as the computational time in the face of a large dataset. Our future work aims at analyzing more application specific attacks to know the patterns in more optimized way.

References

1. Battista Biggio, Member,Giorgio Fumera (2014). Security Evaluation of Pattern Classifiers under Attack, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 26, NO. 4.
2. Olaleye Oludare, Olabiyisi Stephen (2014). An Optimized Feature Selection Technique For Email Classification, INTERNATIONAL JOURNAL OF SCIENTIFIC & TECHNOLOGY RESEARCH VOLUME 3, ISSUE 10, OCTOBER 2014
3. 5 Cristianini, N., & Shawe-Taylor, J. (2000). A introduction to support vector machines and other kernel-based learning methods. Cambridge, UK: Cambridge University Press.
4. Priyanka, C., Rajesh, W., & Sanyam, S. (2010). Spam Filtering using Support Vector Machine||, Special 20 Issue of IJCCT **1**(2, 3, 4), 166-171.
5. Duda, R. O., Hart, P. E., and Stork, D. G., 2001. *Pattern Classification* (2nd ed.). John Wiley and Sons, University of Michigan.
6. Jain, A. K., Duin, R. P. W., Mao, J., 2000. Statistical pattern recognition: a review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22 (1), pp. 4-38.
7. R. Parimala, R. Nallaswamy (2011).A STUDY ON ENHANCING CLASSIFICATION ACCURACY OF SPAM E-MAIL DATASET. International Journal of Computer Trends and Technology- volume2Issue2
8. Robert E. Schapire, "The boosting approach to machine learning: An overview", In D. D. Denison, M. H. Hansen, C. Holmes, B. Mallick, B.Yu, editors, Nonlinear Estimation and Classification, Springer, 2003.

AUTHOR(S) PROFILE

S. Surendar, is currently PG scholar pursuing M.Tech Computer Science and Engineering in SRM University, TamilNadu, India. He received the B.Tech degree in Information Technology from Anna University, TamilNadu, India, in 2010. He has 4 years of experience in development of Data Warehouse applications. His research comforts are Data Warehousing, Data Mining and Cloud Computing.

R. Veeramani, received the B.E degree in Computer science and engineering from University of Madras, TamilNadu, India, in 1998 and M.E in Computer science and engineering from Annamalai University, TamilNadu, India in 2006.He is currently an assistant professor at SRM University, Chennai. He has more than 10 years' experience in teaching and development of e-commerce applications. His research interests are in Data mining, distributed computing, cloud computing, parallel system and Programming languages. In addition, he is member of Indian science congress association.