

# A Robust One-Class Bayesian Approach for Masquerade Detection

Qifan Wang  
Computer Science Department  
Purdue University  
West Lafayette, IN 47907, US  
wang868@cs.purdue.edu

Luo Si  
Computer Science Department  
Purdue University  
West Lafayette, IN 47907, US  
lsi@cs.purdue.edu

## ABSTRACT

Masquerade attack is a serious computer security problem, which can cause significant damage. Many previous research works were based on two-class training that collected data from multiple users to train one self (i.e., regular) model and one non-self (i.e., abnormal) model for each user. Two-class learning methods for masquerade detection can generate accurate results but demand data from all users, which may not be available for many practical applications. On the other side, one-class learning methods build a model for each user by utilizing only his/her own data. One-class learning methods are more practical but they also suffer from the limited amount of training information from a single user. To address the data sparsity issue, we propose a robust one-class Bayesian approach for masquerade detection. The new method explicitly considers model uncertainty by integrating out the unknown model parameters for generating robust results, while previous one-class methods only use a single point estimate to find an optimal model. We derive the full analytical solution of the predictive distribution over all possible model parameters. A set of experimental results demonstrate that the proposed approach outperforms most previous one-class approach for masquerade detection.

## Categories and Subject Descriptors

I.2.6 [Artificial Intelligence]: Learning

## General Terms

Algorithms, Security, Theory

## Keywords

Masquerade Detection, One-Class Learning, Bayesian Approach, Multinomial Distribution

## 1. INTRODUCTION

Masquerade attack is one of the most serious and dangerous security problems, which represents a class of insider or outsider attacks that can occur in many different ways. Detecting masquerade attacks is considered challenging due to two aspects. Firstly, if a masquerader can mimic the user's behavior similarly and successfully, he is not likely to be detected. On the other hand, if the user himself has to

Table 1: Results of previous classification algorithms on the Schonlau Data Set

Algorithm	Hits	False Positives
N.Bayes(updating)	61.5%	1.5%
N.Bayes(no Upd.)	66.2%	4.6%
Uniqueness	39.4%	1.4%
1-Step Markov	69.3%	6.7%
Hybrid Markov	49.3%	3.2%
IPAM	41.4%	2.7%
Sequence-Matching	36.8%	3.7%
Compression	34.2%	5.0%
Semi-Global Alignment	75.8%	7.7%
ECM	72.3%	2.5%
Oc-Naive Bayes(no Upd.)	63.2%	4.7%
Oc-SVM	72.7%	6.8%
<b>Oc-Bayesian</b>	65.3%	4.2%

behaving differently from his trained profile due to some specific reason, the detecting system may misclassify his legal access as a security breach and hence cause false alarms.

Recently, one-class learning approaches have been proposed to address this problem. They treated the task of profiling a user by modeling his own data exclusively without using samples from other users, while achieving good performance and minimal false positive rates.

Although one-class training methods perform equally well as two-class training approaches and are more efficient in terms of training process and testing, they only train one single model, with fixed optimal parameters, from the user's samples. This may lead to imprecise predicting result especially when training data samples from the user is small. In this paper, we propose a robust one-class bayesian approach which takes account of model uncertainty by integrating out the unknown model parameters, while previous methods use a single point estimate to find an optimal model by maximizing the posterior. We also derive the full analytical solution of the predictive distribution over all possible model parameters. Table 1 and serve as a baseline for comparison.

## 2. ONE-CLASS BAYESIAN APPROACH

The problem of small data samples and resulting parameter uncertainty suggests the use of Bayesian techniques. Such an approach offers a natural and principled way to take account of uncertainty by integrating out the unknown model parameters.

In the previous works, such as Oc-Naive Bayes and Oc-

SVM, they both aimed to find a single point estimate for the model parameter vector  $\theta$ . A posterior distribution over  $\theta$ ,  $P(\theta|d)$ , is obtained by combining a prior distribution  $P(\theta)$  with the observation likelihood  $P(d|\theta)$  using Bayes' rule. When the training examples  $d$  are large, we may expect the posterior  $P(\theta|d)$  to reflect this and to be relatively narrow. In the case of a small dataset, the posterior tend to be much broader. Therefore, it would bring the uncertainty of the samples in the value of  $\theta$ .

Although the mode of the posterior could be achieved by existing max-posterior methods, a more powerful approach is to take account of posterior uncertainty when evaluating the probability of a test block  $q$  by computing the predictive distribution:

$$\begin{aligned} P(q|d) &= \int_{\theta} P(q|\theta)P(\theta|d)d\theta \\ &= \int_{\theta} P(q|\theta) \frac{P(d|\theta)P(\theta)}{P(d)} d\theta \\ &= \int_{\theta} P(q|\theta) \frac{P(\theta)}{P(d)} \prod_{j=1}^n P(d^j|\theta)d\theta \end{aligned} \quad (1)$$

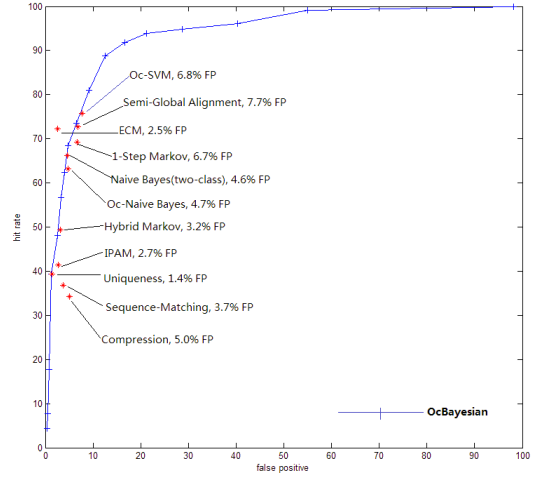
We use the fact that  $d$  and  $q$  are assumed to be generated by the same distribution. From the above equation we can find that the predictive distribution is actually obtained by averaging the probability of  $P(q|\theta)$  under the model over all possible parameter values, weighted by their posterior probability  $P(\theta|d)$ . When the training dataset is large enough, the posterior  $P(\theta|d)$  is peaked around the optimal value  $\theta^*$  then it can be seen that  $P(q|d) \approx P(q|\theta^*)$  and the conventional maximum likelihood predictor is recovered. On the other side, when the posterior is broad, the averaging process accounts for the uncertainty in the model parameter values. In this case, the natural conjugate prior of a *multinomial* distribution is the *Dirichlet* distribution:

$$P(\theta) = Z_{\alpha} \prod_{i=1}^m (\theta_i)^{\alpha_i - 1} \quad (2)$$

where  $\alpha_i$  here is the *hyper parameter* and  $Z_{\alpha} = \frac{\Gamma(\alpha)}{\prod_{i=1}^m \Gamma(\alpha_i)}$  is the normalization constant which does not depend on the parameters  $\theta$ . Under this prior we can compute the resulting posterior, which is also a *Dirichlet* distribution.

$$P(\theta|d) = \frac{\Gamma(|d| + \alpha)}{\prod_{i=1}^m \Gamma(d_i + \alpha_i)} \prod_{i=1}^m (\theta_i)^{d_i + \alpha_i - 1} \quad (3)$$

where  $d_i$  here is the number of times that command  $c_i$  appears in all the training command blocks,  $d_i = \sum_{j=1}^n d_i^j$ , and  $|d| = \sum_i d_i$ . We choose eqn.?? as our prior distribution. This distribution has a number of *hyper-parameters*  $\alpha_i$  equal to the number of parameters in the model which can be interpreted as additional data or pseudo-counts.



**Figure 1: ROC curve for one-class Bayesian model. The best outcomes from other algorithms are also included for comparison.**

We can derive the predictive distribution as follows:

$$\begin{aligned} P(q|d) &= Z_q Z_{d+\alpha} \int_{\theta} \prod_{i=1}^m (\theta_i)^{d_i + \alpha_i - 1} d\theta \\ &= Z_q \frac{\Gamma(|d| + \alpha)}{\prod_{i=1}^m \Gamma(d_i + \alpha_i)} \frac{\prod_{i=1}^m \Gamma(q_i + d_i + \alpha_i)}{\Gamma(|q| + |d| + \alpha)} \\ &= Z_q \frac{\Gamma(|d| + \alpha)}{\Gamma(|q| + |d| + \alpha)} \left[ \prod_i \frac{\Gamma(q_i + d_i + \alpha_i)}{\Gamma(d_i + \alpha_i)} \right] \\ &= Z_q \frac{\prod_i \prod_{k=1}^{q_i} (d_i + \alpha_i + k - 1)}{\prod_{j=1}^{|q|} (|d| + \alpha + j - 1)} \end{aligned} \quad (4)$$

Note that in the last line of the above equation we use the property:  $\Gamma(z+1) = z\Gamma(z)$  and  $\Gamma$  here is the *Gamma function*. This constitutes our new block probability function.

### 3. EVALUATION AND CONCLUSION

Figure 1 shows the ROC curve of our Bayesian approach as well as the best outcomes from other methods. From this figure we can tell that the hit rate of our Bayesian approach is larger than the hit rate of any other method when their false positive rates are same, except for the ECM algorithm which is computationally intensive since it considered the high order information in the data.

This paper presents a robust one-class Bayesian Approach for the masquerade detection problem. The proposed new research builds model for each user by considering model uncertainty, while previous methods used a single point estimate for building each user model. A full analytical solution is derived in the new Bayesian approach by integrating out unknown model parameters. Experimental results have been provided to show that the proposed approach outperforms most previous one-class learning algorithms. The advantage of the proposed new research is more substantial with a small amount of training data.