E-Mail Worm Detection Using Data Mining

Mohammad M. Masud, The University of Texas at Dallas, USA
Latifur Khan, The University of Texas at Dallas, USA
Bhavani Thuraisingham, The University of Texas at Dallas, USA

ABSTRACT

This work applies data mining techniques to detect e-mail worms. E-mail messages contain a number of different features such as the total number of words in message body/subject, presence/absence of binary attachments, type of attachments, and so on. The goal is to obtain an efficient classification model based on these features. The solution consists of several steps. First, the number of features is reduced using two different approaches: feature-selection and dimension-reduction. This step is necessary to reduce noise and redundancy from the data. The feature-selection technique is called Two-phase Selection (TPS), which is a novel combination of decision tree and greedy selection algorithm. The dimension-reduction is performed by Principal Component Analysis. Second, the reduced data is used to train a classifier. Different classification techniques have been used, such as Support Vector Machine (SVM), Naïve Bayes, and their combination. Finally, the trained classifiers are tested on a dataset containing both known and unknown types of worms. These results have been compared with published results. It is found that the proposed TPS selection along with SVM classification achieves the best accuracy in detecting both known and unknown types of worms.

Keywords: e-mail worm; feature selection; Naïve Bayes; principal component analysis; support vector machine

INTRODUCTION

E-mail worm spreads through infected e-mail messages. The worm may be carried by attachment, or the e-mail may contain links to an infected Web site. When the user opens the attachment, or clicks the link, the host gets infected immediately. The worm exploits the vulnerable e-mail software in the host machine to send infected e-mails to addresses stored in address book. Thus, new machines get infected. Worms bring damage to computer and people in various ways. They may clog the network traffic, cause damage to the system, and make the system unstable or even unusable.

The traditional way of worm detection is signature based. A signature is a unique pattern in the worm body that can identify it as a particular type of worm. Thus, a worm can be
detected from its signature. But the problem
with this approach is that it involves significant
amount of human intervention and may take
a long time (from days to weeks) to discover
the signature. Thus, this approach is not useful
against “zero-day” attacks of computer worm.
Besides, signature matching is not effective
against polymorphism.

Thus, there is a growing need for a fast
and effective detection mechanism that requires
no manual intervention. Our work is directed
towards automatic and efficient detection of
e-mail worms. We apply a feature-based ap-
proach for this purpose. A number of features of
e-mail messages have been identified in Martin,
Sewani, Nelson, Chen, and Joseph (2005a) and
discussed in the Feature Reduction and Classi-
fication section. The total number of features is
large, some of which may be redundant or noisy.
So we apply two different feature-reduction
techniques: a dimension-reduction technique
called Principal Component Analysis (PCA)
and our novel feature-selection technique
called Two-phase Selection (TPS) that applies
decision tree and greedy elimination. These
features are used to train a classifier to obtain
a classification model. We use three different
classifiers for this task: Support Vector Machine
(SVM), Naïve Bayes (NB), and a combination
of SVM and NB, mentioned henceforth as the
Series classifier. The Series approach was first
proposed by Martin, Sewani, Nelson, Chen,
and Joseph (2005b).

We use the dataset of (Martin et al., 2005a)
for evaluation purposes. The original data
distribution was unbalanced, so we balance
it by rearranging. We divide the dataset into
two disjoint subsets: the known worms set
or K-Set and the novel worms set or N-Set. The
K-Set contains some clean e-mails and e-mails
infected by five different types of worms. The
K-Set contains e-mails infected by a sixth type
worm, but no clean e-mails. We run a three-fold
cross validation on K-Set and the average accuracy
of novel worm detection on N-Set.

Our contributions to this research work are
as follows: First, we apply two special feature-
reduction techniques to remove redundancy
and noise from data. One technique is PCA, and
the other is our novel TPS algorithm. PCA is
commonly used to extract patterns from high
dimensional data, especially when the data are
noisy. Besides, it is a simple and nonparametric
method. TPS applies decision tree C4.5 (Quin-
lan, 1993) for initial selection, and thereafter,
it applies the greedy elimination technique (see
the Two-phase Feature Selection section). Sec-
ond, we create a balanced dataset as explained
above. Finally, we compare the individual
performances among NB, SVM, and Series
and show empirically that the Series approach
proposed by Martin et al. (2005b) performs
worse than either NB or SVM.

The rest of this article is organized as
follows: The Related Work section describes
related work in automatic e-mail worm detec-
tion; the Feature Reduction and Classification
Techniques section describes the feature-select-
tion, dimension-reduction, and classification
techniques; the Dataset section describes the
distribution of the dataset; the Experimental
Setup section describes the experimental
setup such as hardware, software, and system
parameters; the Results section discusses the
results; and the Conclusion section offers future
guidelines for research.

RELATED WORK
There are different approaches to automate
the detection of worms. These approaches are
mainly of two types: behavioral and content-
based. Behavioral approaches analyze the
behavior of messages like source-destination
addresses, attachment types, message frequency.
Content-based approaches look into the content
of the message and try to detect signature au-
tomatically. There are also combined methods
that take advantage of both techniques.

An example of behavioral detection is
social network analysis (Golbeck & Hendler,
Preserving Privacy in Mining Quantitative Associations Rules
[www.igi-global.com/article/preserving-privacy-mining-quantitative-associations/40357?camid=4v1a](http://www.igi-global.com/article/preserving-privacy-mining-quantitative-associations/40357?camid=4v1a)