Università degli studi di Modena e Reggio Emilia

Dipartimento di Ingegneria

Corso di Laurea Magistrale in Ingegneria Informatica

Adversarial Machine Learning per il Rilevamento di Botnet

Relatore:
Prof. Michele Colajanni

Candidato:
Alessandro Aleotti

Correlatore:
Ing. Mirco Marchetti

Anno Accademico 2017/2018

Indice

Capitolo 1

Introduzione

In questo capitolo si propongono degli esempi per gli oggetti utilizzati più di frequente in latex: la Sezione 1.1 descrive come scrivere citazioni, la Sezione 1.2 propone degli esempi di oggetti float, la Sezione 1.3 descrive come compilare questo documento.

1.1 Citazioni

Inserisco qualche citazione per mostrare la bibliografia. Per gli articoli accademici è quasi sempre possibile reperire i blocchi da inserire nel file bib da scholar, come ad esempio. Scholar in questo caso è una risorsa/sito online e per questo. Precediamo le citazione da uno spazio indivisibile tramite il carattere ~.

1.2 Oggetti float

Nella Sezione 1.2.1 si propone un esempio di figura float, mentre nella Sezione 1.2.2 si propone un esempio di tabella float.

1.2.1 Figure

La Figura 1.1 è un esempio di figura float.

Figura 1.1: Esempio di figura float in latex.

1.2.2 Tabelle

La Tabella 1.1 è un esempio di tabella.

allineamento centrale	allineamento a sinistra	allineamento a destra
centrale	sinistra	destra

Tabella 1.1: Esempio di tabella float in latex.

1.3 Compilazione

Di seguito il codice da utilizzare per generare il pdf:

```
$ pdflatex main.tex
$ bibtex main.aux
$ pdflatex main.tex
$ pdflatex main.tex
```

Capitolo 2

Stato dell'arte

In questo capitolo si propongono degli esempi per gli oggetti utilizzati più di frequente in latex: la Sezione 1.1 descrive come scrivere citazioni, la Sezione 1.2 propone degli esempi di oggetti float, la Sezione 1.3 descrive come compilare questo documento.

Capitolo 3

Progetto

In questo capitolo si propone il progetto realizzato per raggiungere gli obiettivi preposti: si è partiti dalla realizzazione di un classificatore basato su *Random Forest* per poi passare ad una versione più elaborata, utilizzando una rete neurale. Il passo successivo ha riguardato la creazione di una *Generative Adversarial Network* a partire da un Autoencoder.

3.1 Classificatore Random Forest

La prima fase di questo studio è stata quella di implementare un classificatore in grado di separare efficacemente domini DGA da domini non malevoli basandosi unicamente sulle caratteristiche linguistiche dei domini: infatti, ad un esame preliminare, i domini DGA presentano caratteristiche ben differenti da semplici frasi o parole che solitamente compongono i domini reali.

Si è scelto di utilizzare Random Forest in quanto ritenuto il più adatto al caso in esame. L'algoritmo è stato inoltre messo a confronto con *Support Vector Machine* e *Naive-Bayes*.

All'interno del classificatore *Random Forest* [**?**], ogni albero dell'insieme è costruito a partire da un campione estratto con sostituzione dal *training set*. In aggiunta, al momento della divisione del nodo durante la costruzione di un albero, la divisione scelta non è più la migliore soluzione tra tutte le *features*. Al suo posto, la divisone che

viene scelta è la migliore divisione all'interno di un *subset* casuale tra tutte le *features*. Come risultato di questa casualità, il *bias* della foresta di solito aumenta leggermente (rispetto al *bias* di un singolo albero non casuale) ma, a causa della media, la sua varianza diminuisce, di solito compensando l'aumento di *bias*, quindi dando un modello generale migliore.

3.1.1 Dataset

I *dataset* di *training* e *testing* sono stati ricavati due fonti differenti: per quel che riguarda i domini reali si è fatto riferimento alla classifica dei domini più visitati al mondo fornita da *Alexa Internet Inc.* [2] , per un totale di 1 milione di siti realmente esistenti; mentre grazie al repository fornito da [1] è stato possibile ottenere un *dataset* esaustivo di esempi *DGA* da diverse famiglie di *malware*.

A partire da tale dataset combinato si è proceduto alla creazione di un classificatore binario che fosse in grado di distinguere domini reali da domini generati algoritmicamente.

Il passo seguente stato creare una serie di *features* che fossero in grado di descrivere le caratteristiche linguistiche dei domini presi in esame. Per raggiungere tale obiettivo si è fatto riferimento a ricerche già esistenti. Di seguito viene illustrato l'insieme di tali *features*:

3.1.2 Features

- Meaningful Characters Ratio. Models the ratio of characters of the string p that comprise a meaningful word. Low values indicate automatic algorithms. Specifically, we split p into n meaningful subwords wi of at least 3 symbols: $|wi| >= 3$, leaving out as few symbols as possible: $R(d) = R(p) = max((sumfromi = 1ton)|wi|)/|p|$. If p = facebook, $R(p) = (|face| + |book|)/8 = 1$, the prefix is fully composed of meaningful words, whereas p = pub03str, $R(p) = (|pub|)/8 = 0.375$.

- n-gram Normality Score: This class of features captures the pronounceability of a domain name. The more permissible the combinations of phonemes, the more

pronounceable a word is. Domains with a low number of such combinations are likely DGA-generated. We calculate this class of features by extracting the n-grams of p, which are the substrings of p of length n 1, 2, 3, and counting their occurrences in the (English) language dictionary. The features are thus parametric to $n : Sn(d) = Sn(p) := ((sumofn-gramtinp)count(t))/(|p|-n+1)$, where count(t) are the occurrences of the n-gram t in the dictionary L

3.1.3 Output

3.2 Classificatore Neurale

3.2.1 Input

3.2.2 Composizione Interna

3.2.3 Output

3.3 Realizzazione Adversarial Learning

3.3.1 Input

3.3.2 Composizione Interna

3.3.3 Output

Capitolo 4

Implementazione

In questo capitolo si propongono degli esempi per gli oggetti utilizzati più di frequente in latex: la Sezione 1.1 descrive come scrivere citazioni, la Sezione 1.2 propone degli esempi di oggetti float, la Sezione 1.3 descrive come compilare questo documento.

4.1 Classificatore Random Forest

4.1.1 Input

4.1.2 Composizione Interna

4.1.3 Output

4.2 Classificatore Neurale

4.2.1 Input

4.2.2 Composizione Interna

4.2.3 Output

4.3 Realizzazione Adversarial Learning

4.3.1 Input

4.3.2 Composizione Interna

4.3.3 Output

Capitolo 5

Risultati

In questo capitolo si propongono degli esempi per gli oggetti utilizzati più di frequente in latex: la Sezione 1.1 descrive come scrivere citazioni, la Sezione 1.2 propone degli esempi di oggetti float, la Sezione 1.3 descrive come compilare questo documento.

Capitolo 6

Conclusioni

In questo capitolo si propongono degli esempi per gli oggetti utilizzati più di frequente in latex: la Sezione 1.1 descrive come scrivere citazioni, la Sezione 1.2 propone degli esempi di oggetti float, la Sezione 1.3 descrive come compilare questo documento.

Bibliografia

[1] Andrey Abakumov. Dga. https://github.com/andrewaeva/DGA.

[2] Amazon. Alexa. https://www.alexa.com/, visited in Sep. 2017.