Darknet Traffic Classification

Uma abordagem para CICDarknet2020

C. M. Mateus, B. R. Paulo Vitor

March 3, 2021

Estrutura

- Introdução
- Metodologia
- Extração de atributos
- Análise
- Pré-Processamento
- Seleção de modelos
  - Decision Tree
  - Random Forest
- Seleção de Características
- Conclusões e resultados

Introdução

Problema

Traffic Classification:

- Classificação de tráfego de rede relacionadas à aplicação ou origem.
- Técnicas difundidas: aprendizado supervisionado, análise de portas ou estatística dos pacotes enviados ou recebidos

Introdução

Motivação

Uso de dados encriptados:

- Uma avaliação revisada de modelos de classificação para dados de tráfego não indexado.
- Contribuição de novos resultados para o problema de Traffic Classification

Trabalhos relacionados

- Usando a disponibilização do *dataset ISCXVPN2016* trazendo o tráfego de redes VPN.
  - Acurácia 75%
  - Define classes de aplicações: *chat*, *email*, *FTP*, *streaming*, *VOIP* e *VPN*.
- Da mesma forma, usando o *dataset ISCXTor2016* trazendo o tráfego de redes Tor, com as mesmas abordagens que o trabalho anterior
- Manipulação da *dataset ISCXVPN2016*, fundamentando sua abordagem em outras 3 técnicas:
  - Port-Classification
  - Deep Packet Inspection (DPI)
  - Inferência Estatística

Dataset

CICDarknet2020

- União dos *datasets ISCXVPN2016* e *ISCXTor2016*
- Além das classes de aplicação usadas nas duas bases anteriores, há a inserção de uma classe relacionada à origem do tráfego, como sendo benigno ou provindo da Darknet.
- 158659 registros

Dataset

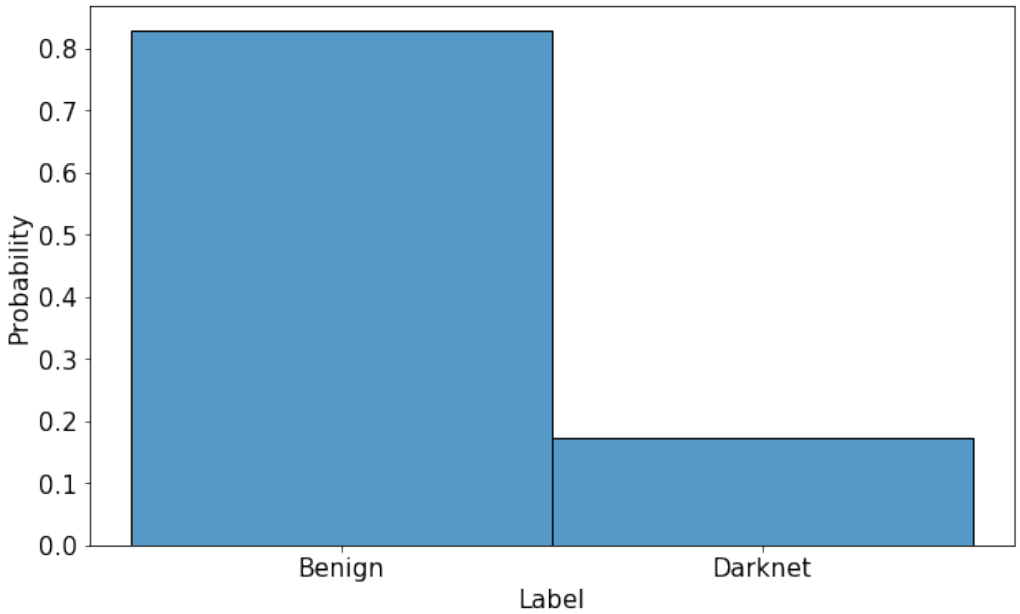

Figure: Relação entre classes de origem dos dados

Dataset

Traffic Category	Applications used
Audio-Stream	Vimeo and Youtube
Browsing	Firefox and Chrome
Chat	ICQ, AIM, Skype, Facebook and Hangouts
Email	SMTPS, POP3S and IMAPS
P2P	uTorrent and Transmission (BitTorrent)
Transfer	Skype, FTP over SSH (SFTP) and FTP over SSL (FTPS) using Filezilla and an external service
Video-Stream	Vimeo and Youtube
VOIP	Facebook, Skype and Hangouts voice calls

Figure: Categorias - CICDarknet2020

Objetivo

Contribuir no aperfeiçoamento da acurácia da caracterização do tráfego usando o novo Dataset, usando modelos simples, tais como Decision Tree e Random Forest.

Metodologia

- Extração de novos atributos relevantes usando os existentes;
- Análise do dataset para verificação da relevância dos atributos
- Pré-processamento da dataset para treinamento dos modelos de classificação
- Avaliação dos modelos treinados
- Seleção de características
- Análise da importância dos atributos para o aprendizado

Extração de atributos

- Criação de novo atributo com a hora do tráfego utilizando o timestamp
- Divisão dos IP's de origem e destino em Unigram, Bigram e Trigram
- Extração de informações extras dos IP's (localização, bogon, etc)

Análise

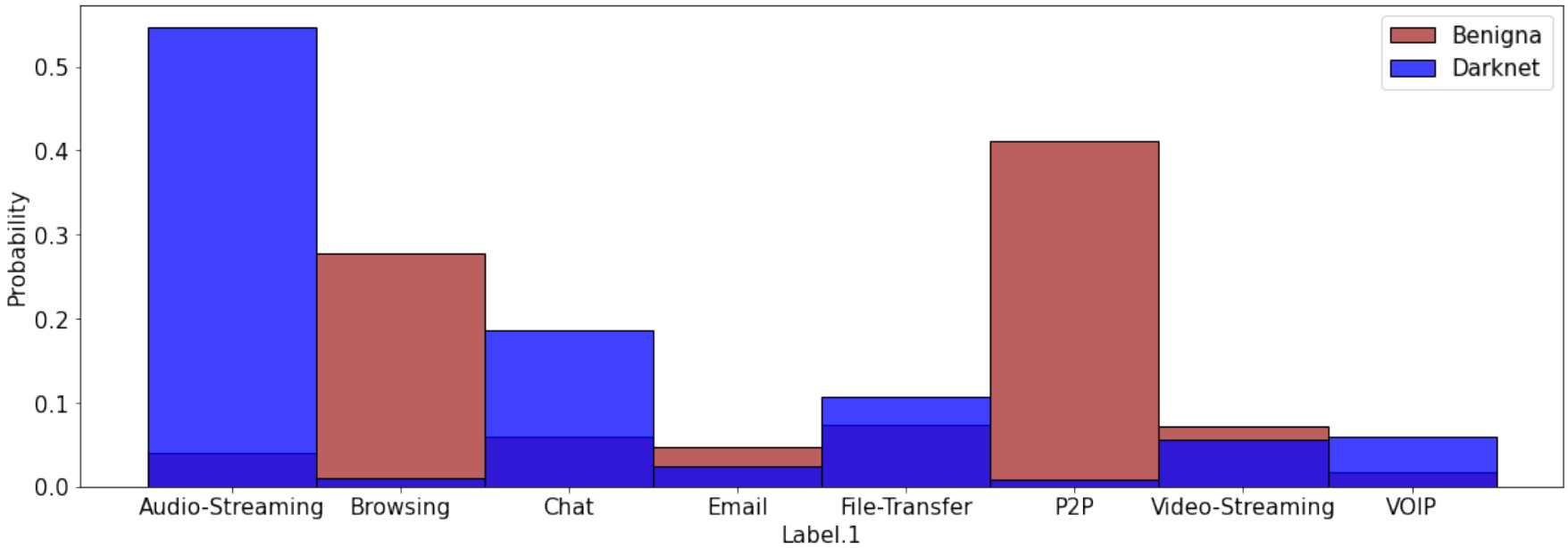

Figure: Relação entre categorias de aplicação

Análise

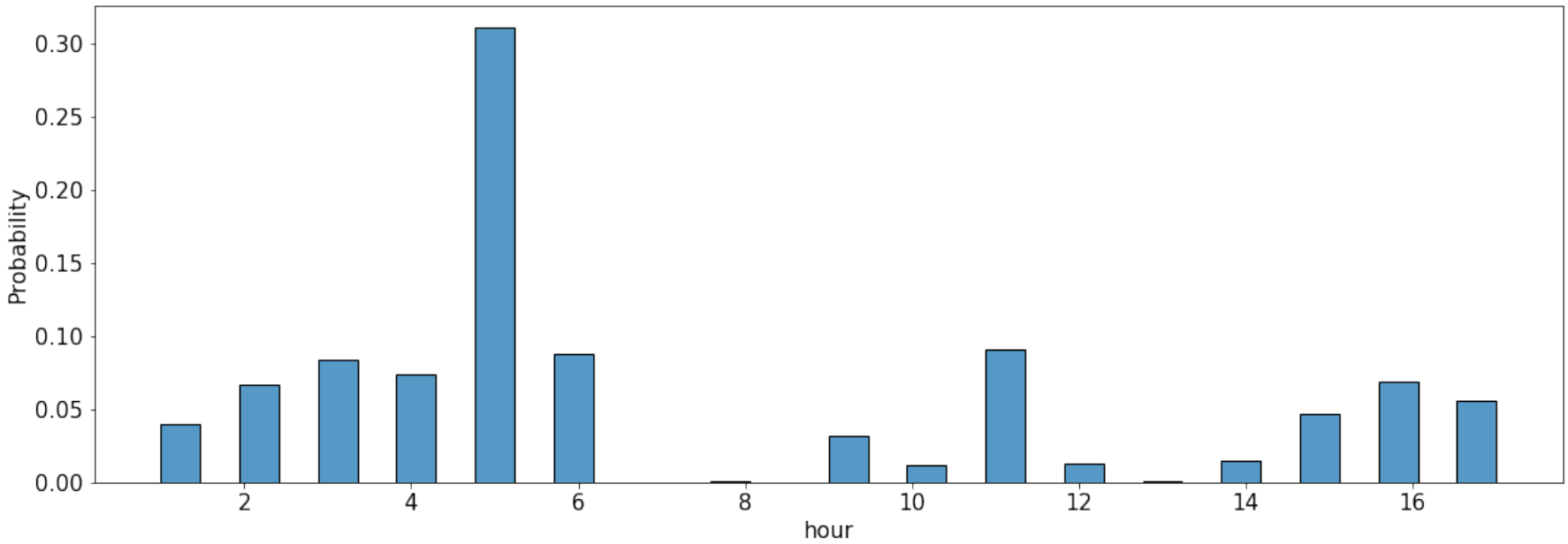

Figure: Horário de Tráfego para redes benignas

Análise

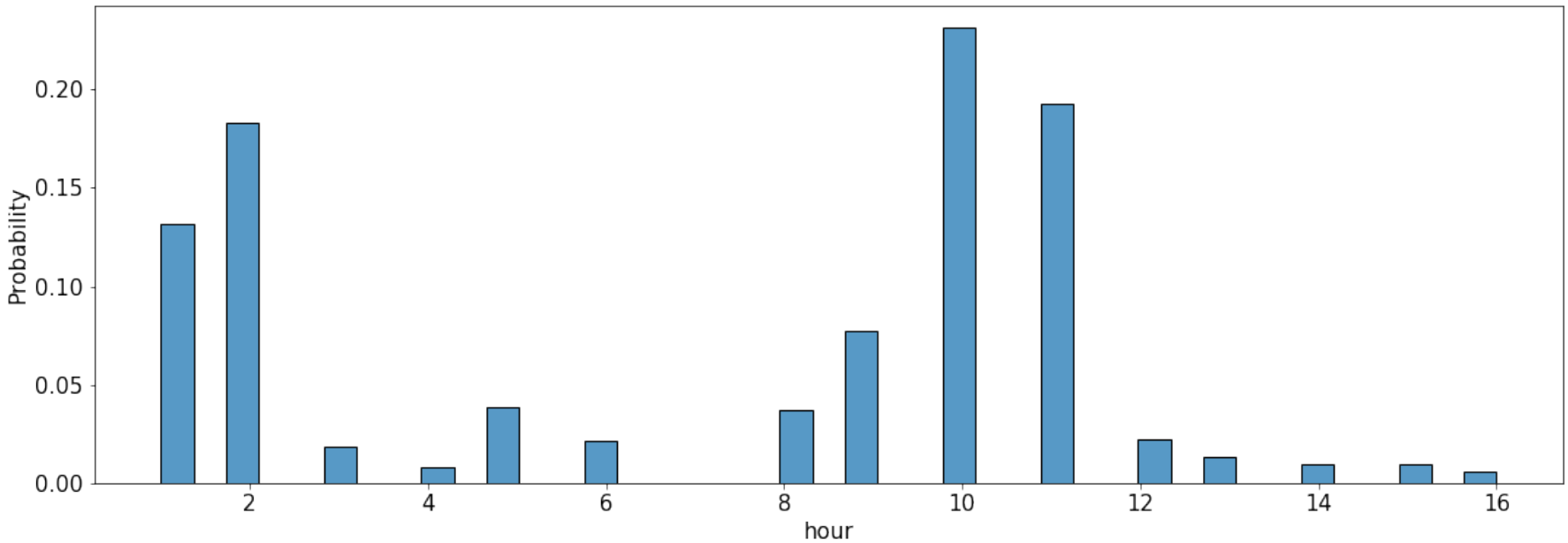

Figure: Horário de Tráfego para redes Darknet

Pré-Processamento

- Correção de Lables redundantes ou fora de padrão
- Remoção de linhas N/A ou com números infinitos
- Escalonamento de atributos numéricos
- Remoção de atributos irrelevantes (FlowID, TimeStamp, IP de origem e destino)
- Hashing Encoding usando os IP's usando tratados (Unigram, Bigram e Trigram)
- Ordinal Encoding (Dados sobre país de origem e destino do IP)

Seleção de modelos

- Dataset dividido para classificação da origem e categorização da aplicação
- Uso dos modelos de Decision Tree e Random Forest
- Avaliação dos modelos pelas métricas de classificação geradas pelo 10-fold estratificado e matriz de confusão computada usando conjunto de teste a parte.
- Seleção de características para definir os atributos mais relevantes usando RFE

Decision Tree - Classificação de origem

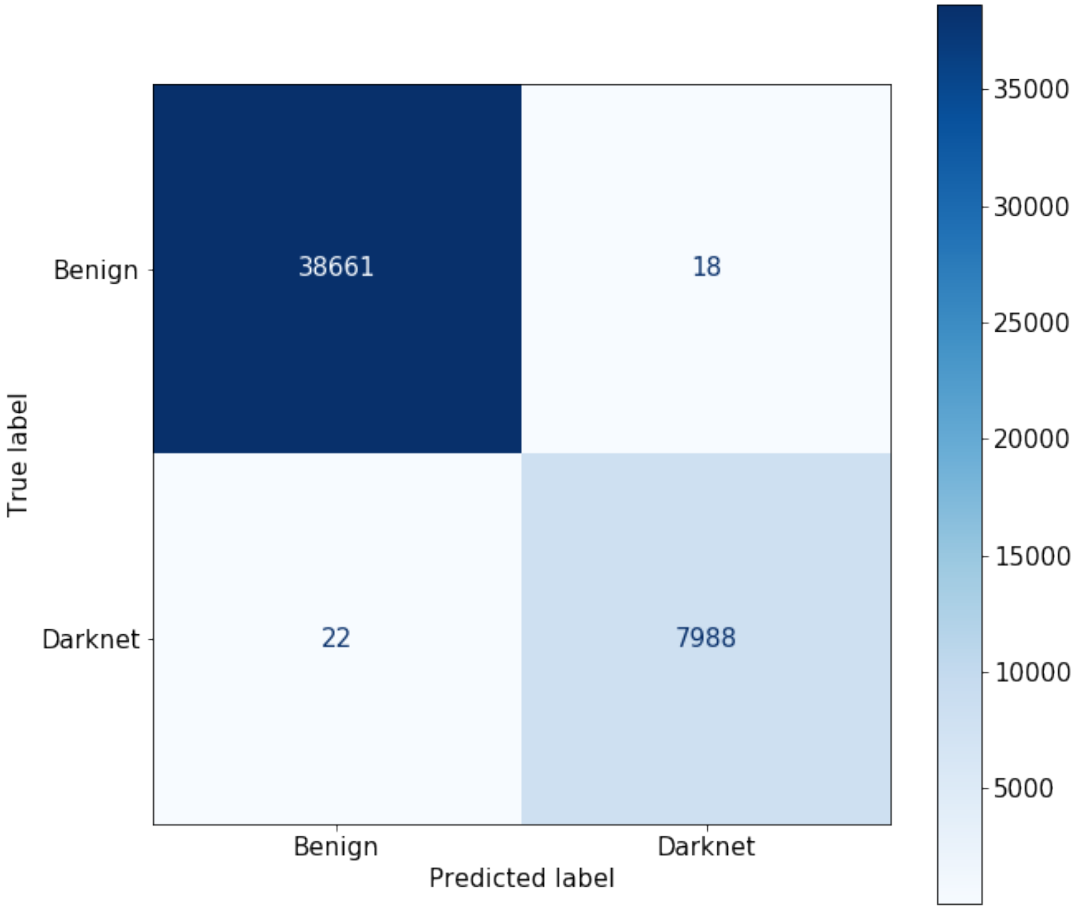

Figure: (Matriz de Confusão) - Classificação origem usando DT

Decision Tree - Classificação de origem

```
              precision    recall  f1-score   support
Benign:       99.94        99.92       99.93       784910.00
Darknet:      99.64        99.71       99.67       163010.00


 10-fold Accuracy: 99.89%
Test accuracy: 99.91%


Confusion matrix:
[[38661    18]
 [   22  7988]]

Benign :    99.94%
Darknet :   99.78%
```

Figure: (Relatório das métricas) - Classificação origem usando DT

Random Forest - Classificação de origem

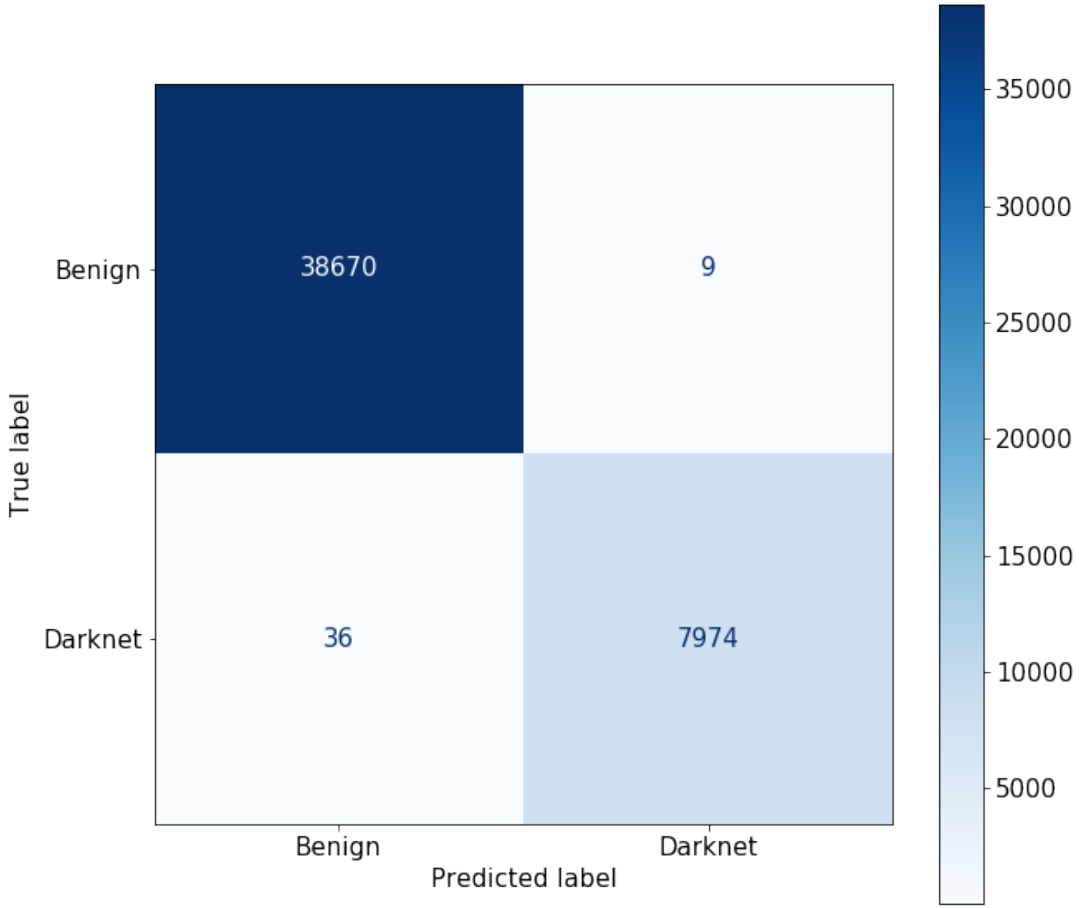

Figure: (Matriz de Confusão - Classificação origem usando RF

Random Forest - Classificação de origem

```
              precision    recall  f1-score   support
Benign:       99.94        99.92       99.93     784910.00
Darknet:      99.64        99.71       99.67     163010.00


 10-fold Accuracy: 99.89%
Test accuracy: 99.91%


Confusion matrix:
[[38661    18]
 [   22  7988]]

Benign :    99.94%
Darknet :   99.78%
```

Figure: (Relatório das métricas) - Classificação origem usando RF

Decision Tree - Categorização da aplicação

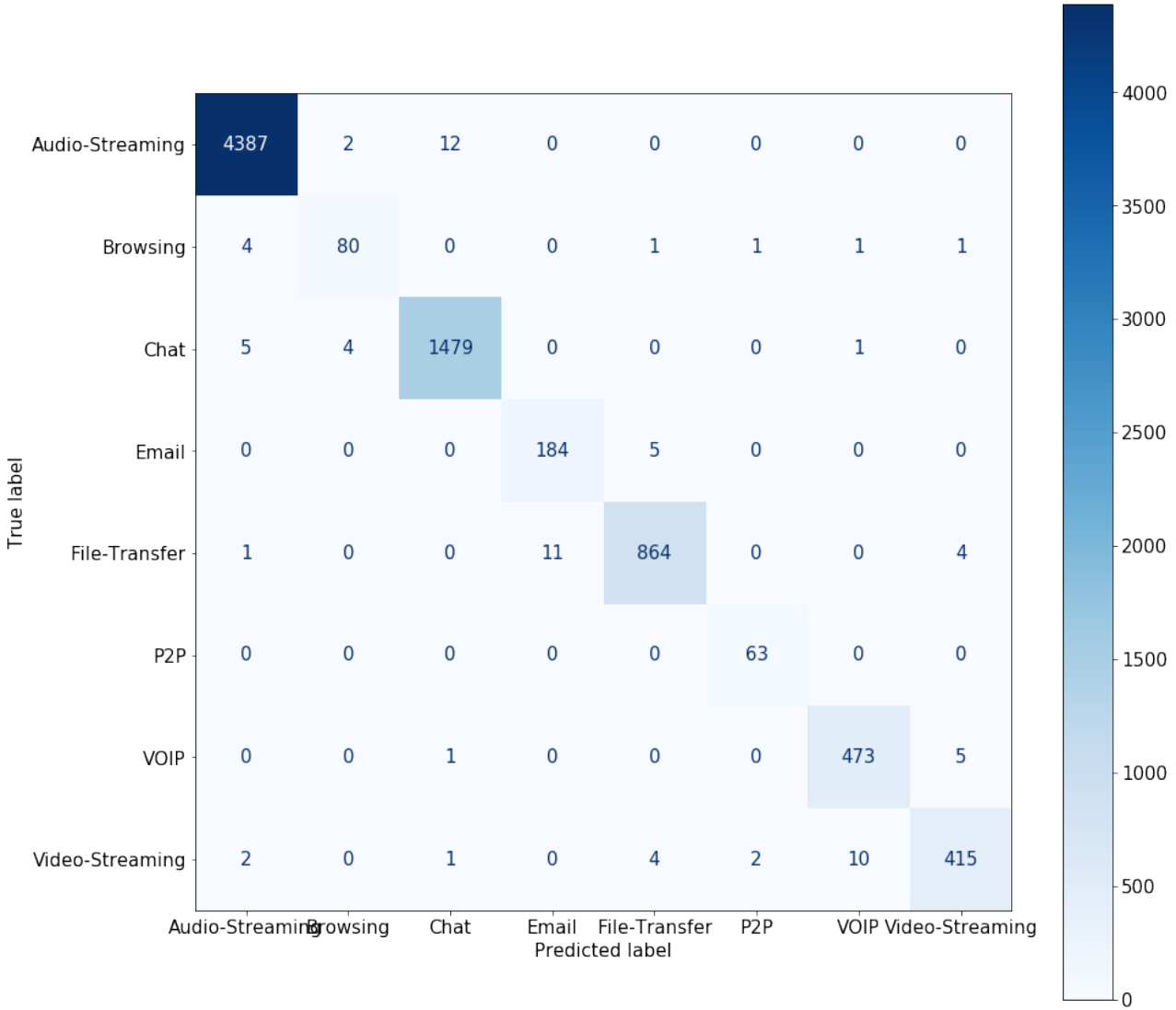

Figure: (Matriz de Confusão) - Categorização da aplicação usando DT

Decision Tree - Categorização da aplicação

```
                precision       recall       f1-score       support
Audio-Streaming:  99.60          99.66           99.63          88830.00
Browsing:         93.05          83.82           87.85          1750.00
Chat:             98.93          99.28           99.10          30520.00
Email:            95.33          95.70           95.45          3930.00
File-Transfer:    98.27          97.92           98.09          17300.00
P2P:              97.12          96.25           96.48          1570.00
VOIP:             97.39          97.97           97.68          9860.00
Video-Streaming:  96.28          96.16           96.21          9120.00


 10-fold Accuracy: 98.81%
Test accuracy: 99.03%


Confusion matrix:
[[4387    2   12    0    0    0    0    0]
 [   4   80    0    0    1    1    1    1]
 [   5    4 1479    0    0    0    1    0]
 [   0    0    0  184    5    0    0    0]
 [   1    0    0   11  864    0    0    4]
 [   0    0    0    0    0   63    0    0]
 [   0    0    1    0    0    0  473    5]
 [   2    0    1    0    4    2   10  415]]

Audio-Streaming :    99.73%
Browsing :           93.02%
Chat :               99.06%
Email :              94.36%
File-Transfer :      98.86%
P2P :                95.45%
VOIP :               97.53%
Video-Streaming :    97.65%
```

Figure: (Relatório das métricas) - Categorização da aplicação usando DT

Random Forest - Categorização da aplicação

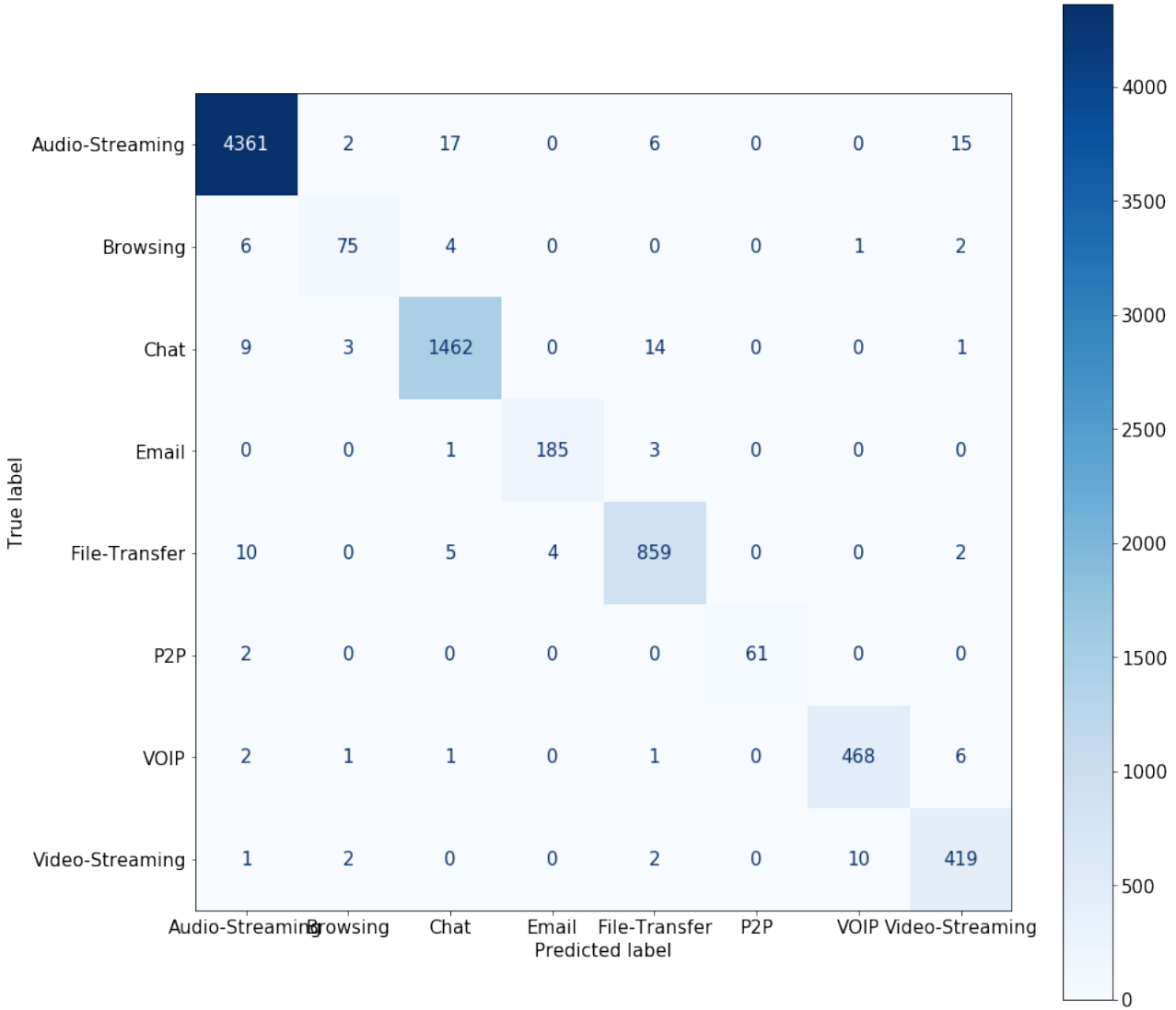

Figure: (Matriz de Confusão - Categorização da aplicação usando RF

Random Forest - Categorização da aplicação

```
                precision       recall       f1-score       support
Audio-Streaming:  99.50          99.08           99.29          88830.00
Browsing:         88.49          83.95           85.87          1750.00
Chat:             98.51          99.08           98.79          30520.00
Email:            98.47          95.94           97.15          3930.00
File-Transfer:    98.03          96.99           97.50          17300.00
P2P:              95.67          97.50           96.32          1570.00
VOIP:             97.13          97.26           97.16          9860.00
Video-Streaming:  91.32          96.27           93.71          9120.00


 10-fold Accuracy: 98.34%
Test accuracy: 98.34%


Confusion matrix:
[[4361    2   17    0    6    0    0   15]
 [   6   75    4    0    0    0    1    2]
 [   9    3 1462    0   14    0    0    1]
 [   0    0    1  185    3    0    0    0]
 [  10    0    5    4  859    0    0    2]
 [   2    0    0    0    0   61    0    0]
 [   2    1    1    0    1    0  468    6]
 [   1    2    0    0    2    0   10  419]]

Audio-Streaming :   99.32%
Browsing :          90.36%
Chat :              98.12%
Email :             97.88%
File-Transfer :     97.06%
P2P :               100.00%
VOIP :              97.70%
Video-Streaming :   94.16%
```

Figure: (Relatório das métricas) - Categorização da aplicação usando RF

Seleção de Características

```
Optimal number of features: 12
Feature                    Importance
-----------------------------------
col_76:                    0.4287
hour:                      0.3350
Idle Max:                  0.0826
Fwd Seg Size Min:          0.0389
col_58:                    0.0326
Flow Duration:             0.0274
Average Packet Size:       0.0161
Src Port:                  0.0117
FWD Init Win Bytes:        0.0090
Flow IAT Std:              0.0078
Dst Port:                  0.0067
col_95:                    0.0035
```

Figure: (Sumário de RFE) - 12 features mais relevantes

Seleção de Características

```
                precision       recall       f1-score       support
Audio-Streaming:  99.72          99.80           99.76          88830.00
Browsing:         96.11          92.48           93.99          1750.00
Chat:             99.51          99.34           99.43          30520.00
Email:            94.03          94.19           94.05          3930.00
File-Transfer:    97.81          97.92           97.86          17300.00
P2P:              98.82          99.38           99.07          1570.00
VOIP:             97.80          97.87           97.82          9860.00
Video-Streaming:  96.77          96.60           96.66          9120.00


 10-fold Accuracy: 99.00%
Test accuracy: 99.13%


Confusion matrix:
[[4388    0   13    0    0    0    0    0]
 [   3   81    1    0    1    1    1    0]
 [   4    4 1481    0    0    0    0    0]
 [   0    0    0  186    3    0    0    0]
 [   0    0    0   10  869    0    0    1]
 [   0    0    0    0    0   63    0    0]
 [   0    0    0    0    0    0  472    7]
 [   1    0    1    0    4    0   15  413]]

Audio-Streaming :   99.82%
Browsing :          95.29%
Chat :              99.00%
Email :             94.90%
File-Transfer :     99.09%
P2P :               98.44%
VOIP :              96.72%
Video-Streaming :   98.10%
```

Figure: (Sumário de RFE) - Avaliação final das features

Conclusões e Resultados

- Acurácia do modelo simples é maior do o reportado pelo artigo de referência
- Contribuição com novas características, atribuindo novos campos ao dataset
- Boa parte dos atributos definidos como relevantes foram constatados como tais

Obrigado!