MINETRAC: Mining Flows for Unsupervised Analysis & Semi-Supervised Classification

P. CASAS, J. MAZEL, **P. OWEZARSKI**

LAAS-CNRS

Toulouse, France

23rd International Teletraffic Congress
ITC 2011

San Francisco, USA

6-8 September 2011

Outline

1. Machine-Learning in TRaffic Analysis & Classification (TRAC)
2. Robust Clustering for Traffic Analysis and Classification
   - Sub-Space Clustering to Improve Robustness
   - Multiple Evidence Accumulation
   - Semi-Supervised Classification
3. Evaluations in Real Traffic Traces
   - The Traffic Datasets
   - SSC-EA Performance vs Traditional Clustering
   - Semi-Supervised Classification Performance
4. Concluding Remarks

Machine-Learning (ML) in TRAC

ML was introduced to enhance port/payload-based traffic classification:

Supervised ML: based on what I ALREADY KNOW

- (+) improves traditional classification techniques.
- (−) needs training on full-labeled traffic datasets.
- (−) labeling traffic flows is difficult, time-consuming, and costly.

Unsupervised ML: KNOWLEDGE-INDEPENDENT analysis

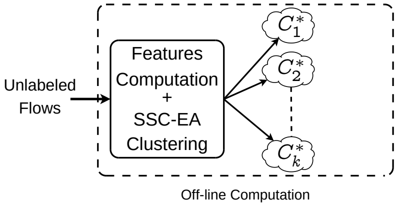
- (+) **Clustering**: separate flows in classes sharing similar characteristics.
- (+) classification is done by limited labeled traffic (**Semi-Supervised ML**).
- (−) lack of robustness: general clustering algorithms are sensitive to initialization, specification of number of clusters, etc.
- (−) difficult to cluster high-dimensional data: structure-masking by irrelevant features, sparse spaces (“the curse of dimensionality”).

Machine Learning in TRAC: our Proposal

We want to reduce the need of labeled traffic, limiting the impacts on classification accuracy.

Machine Learning in TRAC: our Proposal

We want to reduce the need of labeled traffic, limiting the impacts on classification accuracy.

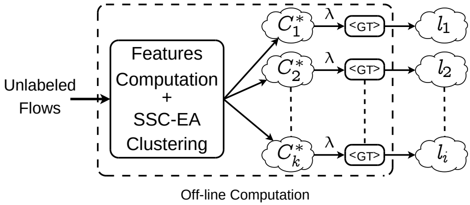


- Two-steps approach: Clustering + Semi-Supervised Classification.

Robust Clustering on unlabeled traffic flows: enhance clustering through the combination of Sub-Space Clustering + Evidence Accumulation.

Machine Learning in TRAC: our Proposal

We want to reduce the need of labeled traffic, limiting the impacts on classification accuracy.



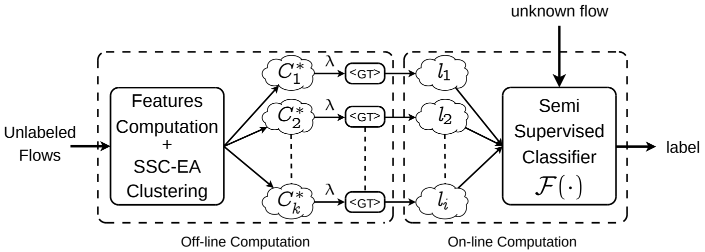
- Two-steps approach: Clustering + Semi-Supervised Classification.

Robust Clustering on unlabeled traffic flows: enhance clustering through the combination of Sub-Space Clustering + Evidence Accumulation.

- Label Clusters: use a small fraction λ of labeled flows per cluster.

Machine Learning in TRAC: our Proposal

We want to reduce the need of labeled traffic, limiting the impacts on classification accuracy.



- Two-steps approach: Clustering + Semi-Supervised Classification.

Robust Clustering on unlabeled traffic flows: enhance clustering through the combination of Sub-Space Clustering + Evidence Accumulation.

- Label Clusters: use a small fraction λ of labeled flows per cluster.
- Distance-based Classification: assign closest-cluster's label.

Outline

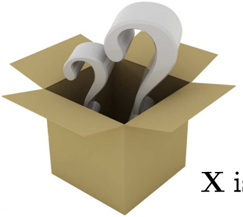
1. Machine-Learning in TRaffic Analysis & Classification (TRAC)
2. Robust Clustering for Traffic Analysis and Classification
   - Sub-Space Clustering to Improve Robustness
   - Multiple Evidence Accumulation
   - Semi-Supervised Classification
3. Evaluations in Real Traffic Traces
   - The Traffic Datasets
   - SSC-EA Performance vs Traditional Clustering
   - Semi-Supervised Classification Performance
4. Concluding Remarks

Clustering for Traffic Analysis (Off-line)

- Let $\mathbf{Y} = \{\mathbf{y}_1, \ldots, \mathbf{y}_n\}$ be a set of n flows captured at the network of analysis.
- Each flow $\mathbf{y}_i \in \mathbf{Y}$ is described by a set of m traffic features: $\mathbf{x}_i = (x_i(1), .., x_i(m)) \in \mathbb{R}^m$.
- $\mathbf{X} = \{\mathbf{x}_1, .., \mathbf{x}_n\}$ is the complete matrix of features, referred to as the *feature space*.

Clustering for Traffic Analysis (Off-line)

- Let $\mathbf{Y} = \{\mathbf{y}_1, \ldots, \mathbf{y}_n\}$ be a set of n flows captured at the network of analysis.
- Each flow $\mathbf{y}_i \in \mathbf{Y}$ is described by a set of m traffic features: $\mathbf{x}_i = (x_i(1), .., x_i(m)) \in \mathbb{R}^m$.
- $\mathbf{X} = \{\mathbf{x}_1, .., \mathbf{x}_n\}$ is the complete matrix of features, referred to as the *feature space*.



$\mathbf{X}$ is a black box

Retrieve natural groupings in $\mathbf{X}$ through clustering is challenging!!!

Outline

1. Machine-Learning in TRaffic Analysis & Classification (TRAC)

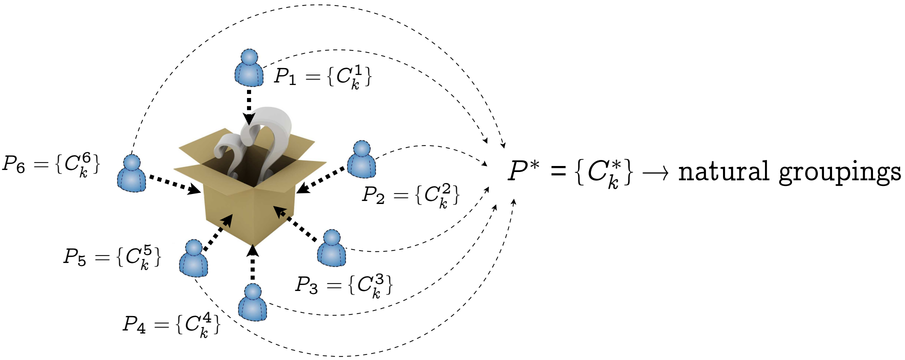
2. Robust Clustering for Traffic Analysis and Classification
   - Sub-Space Clustering to Improve Robustness
   - Multiple Evidence Accumulation
   - Semi-Supervised Classification

3. Evaluations in Real Traffic Traces
   - The Traffic Datasets
   - SSC-EA Performance vs Traditional Clustering
   - Semi-Supervised Classification Performance

4. Concluding Remarks

How to Improve Clustering Robustness?

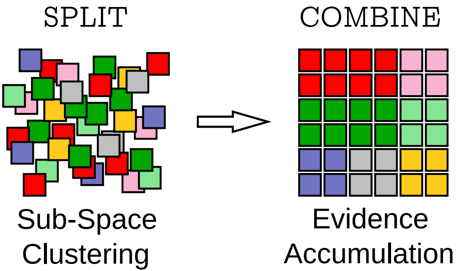
- Idea: combine the information provided by multiple partitions of $\mathbf{X}$ to "filter noise", easing the discovery of natural groupings.
- How to produce multiple partitions? $\longrightarrow$ Sub-Space Clustering.
- Each sub-space $\mathbf{X}_i \subset \mathbf{X}$ is obtained by projecting $\mathbf{X}$ in k out of the m original dimensions. Density-based clustering (DBSCAN) at $\mathbf{X}_i$.



Outline

1. Machine-Learning in TRaffic Analysis & Classification (TRAC)
2. Robust Clustering for Traffic Analysis and Classification
   - Sub-Space Clustering to Improve Robustness
   - Multiple Evidence Accumulation
   - Semi-Supervised Classification
3. Evaluations in Real Traffic Traces
   - The Traffic Datasets
   - SSC-EA Performance vs Traditional Clustering
   - Semi-Supervised Classification Performance
4. Concluding Remarks

Evidence Accumulation to Retrieve Natural Groupings

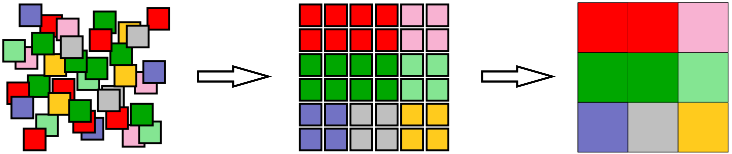


Using Sub-Space Clustering we have SPLIT the problem, how do we COMBINE the obtained partitions? $\longrightarrow$ Evidence Accumulation

- Build a new inter-flows similarity measure S from the N partitions P_i.
- Flows belonging to a natural cluster C_k^* are likely to be co-located in the same cluster in different partitions P_i at different sub-spaces $\mathbf{X}_i$.
- $S(i,j) = n_{ij}/N$, where n_{ij} is the $\#$ of times that flows $\mathbf{y}_i$ and $\mathbf{y}_j$ were assigned to the same cluster through the N partitions.

Evidence Accumulation to Retrieve Natural Groupings

SPLIT　　COMBINE　　MERGE



Sub-Space Clustering　　Evidence Accumulation　　Hierarchical Clustering

Using Sub-Space Clustering we have SPLIT the problem, how do we COMBINE the obtained partitions? $\longrightarrow$ Evidence Accumulation

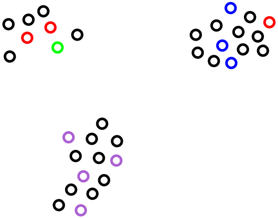
The final partition $P^* = \{C_k^*\}$ is obtained by Hierarchical Clustering on S, MERGING the most similar flows into clusters C_k^*.

Outline

1. Machine-Learning in TRaffic Analysis & Classification (TRAC)
2. Robust Clustering for Traffic Analysis and Classification
   - Sub-Space Clustering to Improve Robustness
   - Multiple Evidence Accumulation
   - Semi-Supervised Classification
3. Evaluations in Real Traffic Traces
   - The Traffic Datasets
   - SSC-EA Performance vs Traditional Clustering
   - Semi-Supervised Classification Performance
4. Concluding Remarks

Semi-Supervised Classification

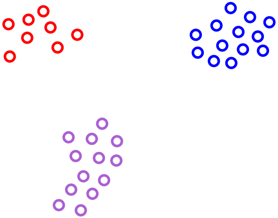
We build a classifier $\mathcal{F}(\cdot)$ from the obtained clusters:



- "Dig" the labels of a small fraction λ of flows (e.g., through DPI).

Semi-Supervised Classification

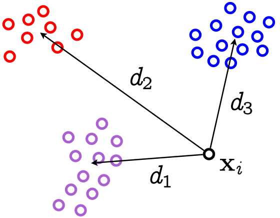
We build a classifier $\mathcal{F}(\cdot)$ from the obtained clusters:



- “Dig” the labels of a small fraction λ of flows (e.g., through DPI).
- Maximum-Likelihood Labeling: label each cluster with the most present label among the λ flows.

Semi-Supervised Classification

We build a classifier $\mathcal{F}(\cdot)$ from the obtained clusters:



- "Dig" the labels of a small fraction λ of flows (e.g., through DPI).
- Maximum-Likelihood Labeling: label each cluster with the most present label among the λ flows.
- Classify an unknown flow $\mathbf{y}_i$ based on its distance to the centroid of each cluster:

$$\mathrm{label}_i = \mathcal{F}(\mathbf{x}_i) = \mathrm{label}\left(\arg\min_k d(\mathbf{x}_i, \mathbf{o}_k^*)\right)$$

Outline

1. Machine-Learning in TRaffic Analysis & Classification (TRAC)
2. Robust Clustering for Traffic Analysis and Classification
   - Sub-Space Clustering to Improve Robustness
   - Multiple Evidence Accumulation
   - Semi-Supervised Classification
3. Evaluations in Real Traffic Traces
   - The Traffic Datasets
   - SSC-EA Performance vs Traditional Clustering
   - Semi-Supervised Classification Performance
4. Concluding Remarks

Outline

1 Machine-Learning in TRaffic Analysis & Classification (TRAC)

2 Robust Clustering for Traffic Analysis and Classification
- Sub-Space Clustering to Improve Robustness
- Multiple Evidence Accumulation
- Semi-Supervised Classification

3 Evaluations in Real Traffic Traces
- The Traffic Datasets
- SSC-EA Performance vs Traditional Clustering
- Semi-Supervised Classification Performance

4 Concluding Remarks

Traffic Datasets and Traffic Features

UNIBIS dataset (2000 flows)

- Controlled campus network traffic, labeled through GT classifier.
- 4 traffic classes: HTTP, eMail (SSL), P2P (BitTorrent, Edonkey), and VoIP (Skype) (500 flows per traffic class).

VALTC dataset (4000 flows)

- Controlled isolated network traffic, labeled through GT classifier.
- 8 traffic classes: HTTP, eMail (POP3), P2P (Emule, LimeWire, Azureus), VoIP (Skype), monitoring traffic, file hosting/download.

Standard 22 Traffic Features

- proto, flow duration, flow volume (bytes and pkts), pkt length (min, mean, max, dev), and inter-arrival time (min, mean, max, dev).
- features are computed in both directions.

Outline

1 Machine-Learning in TRaffic Analysis & Classification (TRAC)

2 Robust Clustering for Traffic Analysis and Classification
- Sub-Space Clustering to Improve Robustness
- Multiple Evidence Accumulation
- Semi-Supervised Classification

3 Evaluations in Real Traffic Traces
- The Traffic Datasets
- SSC-EA Performance vs Traditional Clustering
- Semi-Supervised Classification Performance

4 Concluding Remarks

SSC-EA vs DBSCAN vs k-means

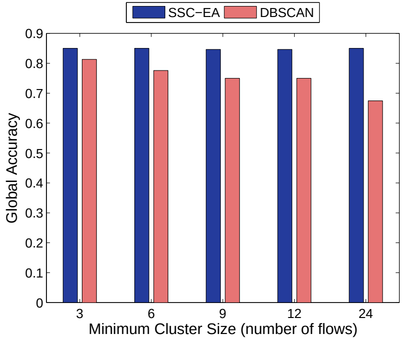
We measure clustering performance through Global Accuracy (GA) and Average per-Cluster Homogeneity (ACH):

$$\boxed{\mathrm{GA} = \frac{\sum_{k=1}^{n_{\text{cls}}} TP(k)}{n}, \quad \mathrm{ACH} = \frac{\sum_{k=1}^{n_{\text{cls}}} \frac{TP(k)}{n(k)}}{n_{\text{cls}}}}$$

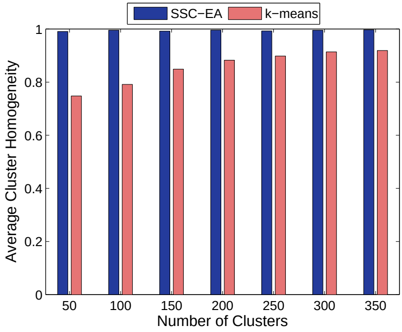
- $TP(k)$: correctly classified flows in cluster k ($\lambda = 100\%$).
- $n(k)$: number of flows in cluster k.
- n_{cls}: number of clusters.

- evaluations performed in UNIBIS.
- SSC-EA vs traditional clustering: DBSCAN and k-means.
- evaluate the impact of Feature Selection (FS) in clustering algorithms.

SSC-EA vs DBSCAN vs k-means



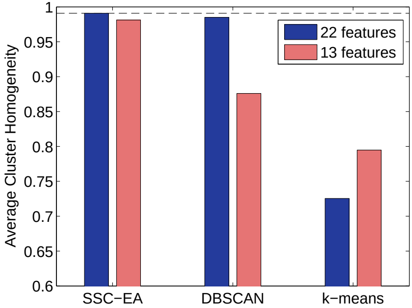
(a) GA: SSC-EA vs DBSCAN.



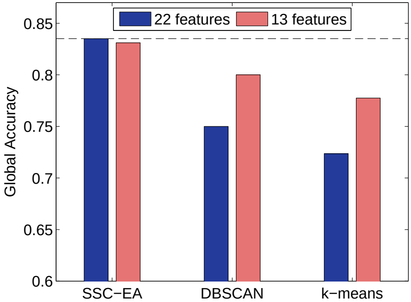
(b) ACH: SSC-EA vs k-means.

- SSC-EA is more robust than DBSCAN regarding clusters' size.
- SSC-EA achieves almost perfect ACH, highly improving k-means.
- SSC-EA GA is about 85%, with about 50 identified clusters.
- SSC-EA GA is impacted by some big-clusters with poor homogeneity.

Impacts of Feature Selection (FS) - Masking Features.



(a) Average per-cluster homogeneity.



(b) Global accuracy.

- GA for the 22 features, and a reduced set of 13 features obtained by FS.
- Selected features correspond mainly to flow volume and packet size features (independent of network conditions).
- SSC-EA is more robust against irrelevant or redundant features.
- The number of SSC-EA clusters falls to about 30 with 13 features.

Outline

1 Machine-Learning in TRaffic Analysis & Classification (TRAC)

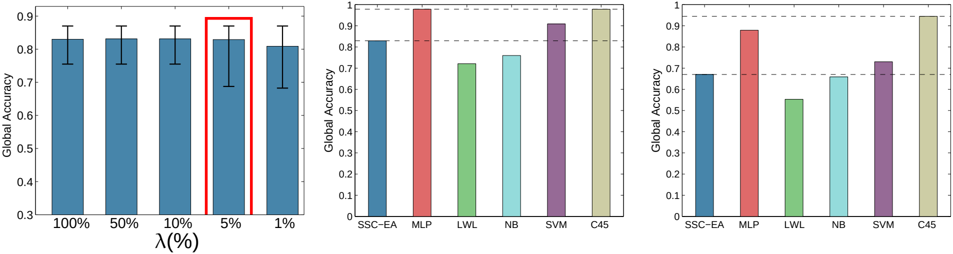
2 Robust Clustering for Traffic Analysis and Classification
- Sub-Space Clustering to Improve Robustness
- Multiple Evidence Accumulation
- Semi-Supervised Classification

3 Evaluations in Real Traffic Traces
- The Traffic Datasets
- SSC-EA Performance vs Traditional Clustering
- Semi-Supervised Classification Performance

4 Concluding Remarks

Semi-Supervised vs Supervised Classification

- The GA of SSC-EA slightly varies with λ (high homogeneity).
- Compare SSC-EA ($\lambda = 5\%$) against “full” supervised classifiers ($\lambda = 100\%$): C45, SVM, Neural Networks (NN), Bayes, and LWL.



(a) $\mathrm{GA}(\lambda)$ (b) GA (UNIBIS) (c) GA (VALTC)

- Difficult to compete with C45, SVM, NN (full training set, $\lambda = 100\%$).
- But limited labeled traffic provides a means for operational deployment.
- Periodically run SSC-EA to recalibrate the limited-reference classifier.

Outline

1. Machine-Learning in TRaffic Analysis & Classification (TRAC)
2. Robust Clustering for Traffic Analysis and Classification
   - Sub-Space Clustering to Improve Robustness
   - Multiple Evidence Accumulation
   - Semi-Supervised Classification
3. Evaluations in Real Traffic Traces
   - The Traffic Datasets
   - SSC-EA Performance vs Traditional Clustering
   - Semi-Supervised Classification Performance
4. Concluding Remarks

Concluding Remarks and Challenges

- Reducing the need of labeled traffic is paramount to achieve useful traffic classifiers.
- Unsupervised analysis based on clustering provides a means to achieve this goal, but robust clustering is difficult to perform.
- SSC-EA improves robustness of analysis by combining multiple outlooks of the same set of flows.
- Feature selection is crucial in any classification problem, and represents a major challenge in an unsupervised context.
- Sub-Space Clustering represents an interesting paradigm for Robust Unsupervised Data Analysis.
- We have applied SSC-EA for Autonomous Network Security with very promising results.

Thank You for Your Attention!!

Remarks & Questions?