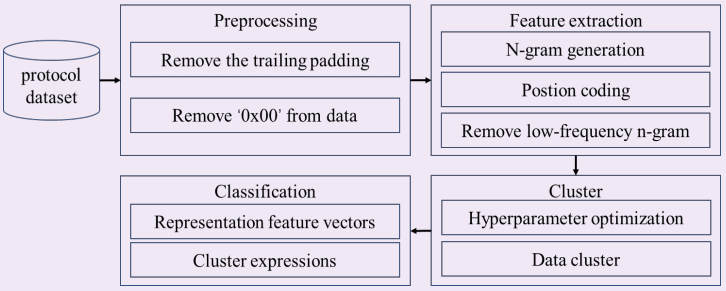
An Unknown Protocol Classification Method based on Clustering Algorithm and Hyperparametric Optimization

HaiYang Liu, Linghang Meng, and Yuantao Gu
Department of Electronic Engineering, Tsinghua University, Beijing, China
Email: {liuhy19, menglh17}@mails.tsinghua.edu.cn, gyt@tsinghua.edu.cn

Background

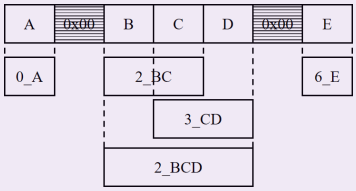
- Unknown Protocol Classification
  - Feature extraction for unknown protocols, classifying and identifying unknown protocols
  - widely applied in network management and maintenance
- Unknown protocol classification method based on clustering algorithm and hyperparametric optimization
  - Applications
    - Unencrypted、Irregular data: direct Subspace clustering suffers from high computational complexity Protocol standards do not apply, fields meaning is not known
- Applications concerns
  - Is the classification method applicable to various protocol data?
    - The method is applicable to non-encrypted data and Encrypted data with certain feature



Motivation: Since standard protocols can be resolved and regulated normally, how can unknown protocols be supervised and utilized?

Main Tasks

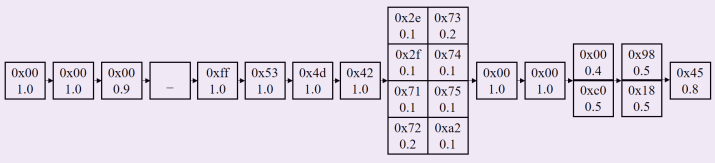
- Feature extraction
  - By introducing the position-coding scheme, n-gram features with semantic information are extracted



- Clustering based on Hyperparametric optimization
  - Combining SC(Silhouette Coefficient), cluster number threshold and noise percentage control, the uniform ϵ-PSO is proposed

$$F_{fitness} = SC - \frac{|cluster_{num} - cluster_{thd}|}{cluster_{thd}} - noise$$

  Where $cluster_{thd}$ and $noise$ are priori value
- Cluster expressions extraction
  - Based on the obtained clustering results, weighted dictionary tree is constructed to obtain cluster expressions



- Constructing Classifiers
  - Two types of classifiers are constructed from clustering results
    - Representative feature vector: Advantages: precision and recall score are both relatively high; Disadvantages: after preprocessing, data have to be mapped to the feature space
    - Cluster expression matching: Advantages: high efficiency, direct matching; Disadvantages: some semantic features are missing, precision and recall score are both relatively low

Simulation

- Hyperparametric optimization

Termination Condition	PSO	ϵ-PSO	Uniform ϵ-PSO
20	0.795	0.823	**0.827**
30	0.802	0.837	**0.845**
50	0.815	0.846	**0.851**

- Cluster result of DBSCAN for Iris

Method	Eps	minPts	ARI	F1_score
DBSCAN	0.436	4	0.522	0.419
I-DBSCAN	0.405	6	0.637	0.641
AF-DBSCAN	0.389	7	0.568	0.582
KANN-DBSCAN	0.434	8	0.612	0.615
Uniform ϵ-PSO	0.401	4	**0.720**	**0.662**

- Comparison of Classifiers