Random Forest of Tensors (RFoT)

Maksim E. Eren, Charles Nicholas, Renee McDonald, and Chris Hamer
(meren1, nicholas)@umbc.edu
Department of Computer Science and Electrical Engineering, UMBC

Objectives

Traditional Machine Learning approaches fail to capture the multi-dimensional details of malware. We introduce a novel methodology to tackle this problem:

- Tensor factorization is a powerful unsupervised learning method.
- Unique **perspectives/patterns are extracted from distinct tensor configurations**.
- Generate forest of random tensor configurations to exploit the **wisdom of crowds philosophy**.
- Clustering can capture the patterns.
- We **vote on the classes in a semi-supervised manner** on the extracted clusters.

CP Decomposition

Build a d dimensional tensor $\mathcal{X}$ shaped $n_1 \times n_2 \times \cdots \times n_d$:

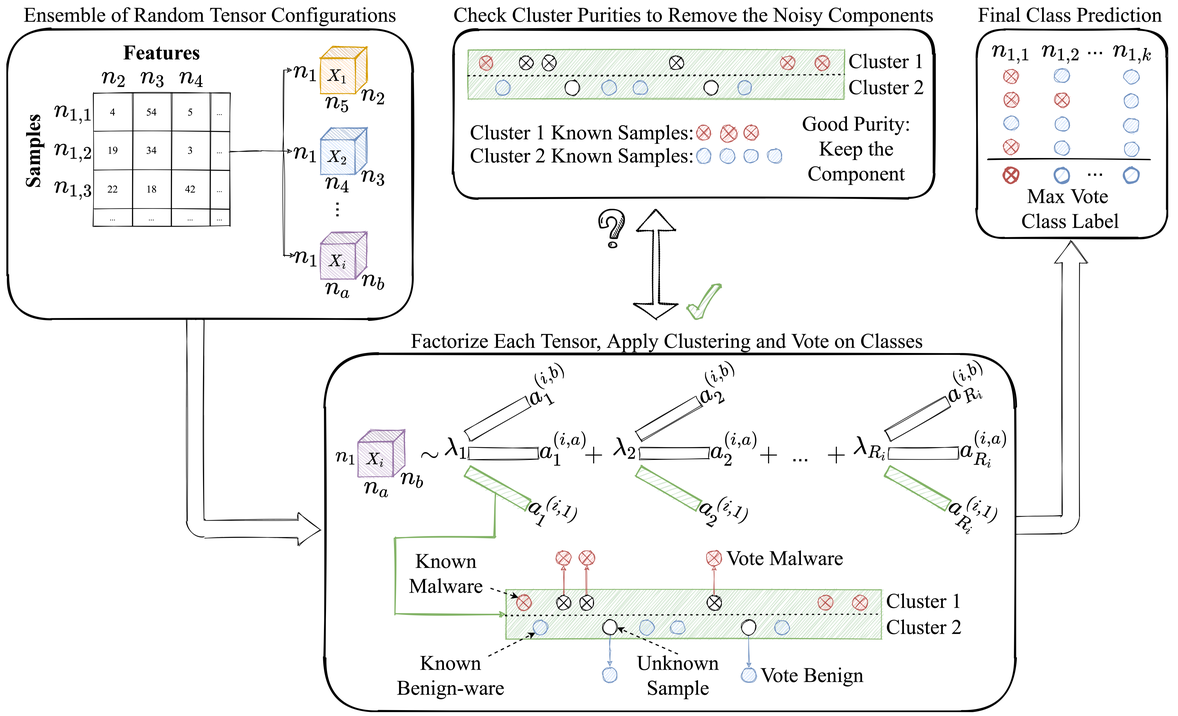
- First dimension represents each of the n_1 specimens in the dataset.
- CP decomposition is written as $\mathcal{X} \approx \mathcal{M} = \lambda\ ;\ \mathbf{A}^{(1)}, \mathbf{A}^{(2)}, \cdots, \mathbf{A}^{(d)}$
- $\mathcal{M}$ is the low-rank approximation of $\mathcal{X}$.
- $\mathbf{A}^{(d)} = [\mathbf{a}_1^{(d)}, \mathbf{a}_2^{(d)}, \ldots, \mathbf{a}_R^{(d)}]$ is the set of R latent factor vectors for dimension d.
- CP-ALS [1, 2, 3] group samples of one class in each of the R factor vectors $\mathbf{a}_r^{(1)} \in \mathbb{R}^{1 \times n_1}$ for the first dimension.

GMM for Capturing the Patterns

Specimens of different classes form groupings in/among the components:

- **Apply the Gaussian Mixture Model (GMM) clustering algorithm to capture patterns.**
- Clustering is applied to each latent factor for the first dimension $\mathbf{a}_r^{(i,1)}$ within each R_i component, for each of the i tensor configurations $\mathcal{X}_i$.

Dataset	Num. Random tensors	F1 Score
EMBER-2018	500	0.9196
IRIS	2000	0.9302
20 Newsgroup	500	0.88



Semi-supervised Ensemble Learner

Known samples are used to vote on specimen class and for identifying and removing bad clusters:

- Utilize the cluster **purity score threshold to remove noisier components.**
- **Each r component of the ith tensor votes on the sample classes over j clusters** $c_{i,r,j}$ using the same known instances.
- Class prediction can be obtained by performing a **majority vote on each sample.**

Experiments and Results

Precise classification results using **only 2% of the corpus in classifying the remaining of the data**:

- EMBER-2018 [4] dataset used to classify malware and benign-ware.
- PE header information in the executables are the features.
- Classification was not possible for the samples where no informative patterns are detected (around 50% of the corpus).
- **.92+ F1 scores are achieved when classification is possible.**
- Works on other datasets; IRIS and 20 Newsgroup.
- **RFoT can help in identifying recent malware with little in the way of labelled data.**

References

Casey Battaglino, G. Ballard, and T. Kolda.
A practical randomized CP tensor decomposition.
SIAM J. Matrix Anal. Appl., 39:876–901, 2018.

Brett W. Bader and Tamara G. Kolda.
Algorithm 862: MATLAB tensor classes for fast algorithm prototyping.
ACM Transactions on Mathematical Software, 32(4):635–653, December 2006.

Tamara G. Kolda and Brett W. Bader.
Tensor decompositions and applications.
SIAM Review, 51(3):455–500, September 2009.

H. S. Anderson and P. Roth.
EMBER: An Open Dataset for Training Static PE Malware Machine Learning Models.
ArXiv e-prints, April 2018.

Presented at the *Malware Technical Exchange Meeting* (MTEM), Online, July 13-15, 2021.