Random Forest of Tensors (RFoT)

Maksim E. Eren*[1], Charles Nicholas†[1], Renee McDonald[1], and Chris Hamer[1]

[1]Department of Computer Science and Electrical Engineering, UMBC
Baltimore, MD, USA 21045

July 14, 2021

1 Abstract

Machine learning has become an invaluable tool in the fight against malware. Traditional supervised and unsupervised methods are not designed to capture the multi-dimensional details that are often present in cyber data. In contrast, tensor factorization is a powerful unsupervised data analysis method for extracting the latent patterns that are hidden in a multi-dimensional corpus. In this poster we explore the application of tensors to classification, and we describe a hybrid model that leverages the strength of multi-dimensional analysis combined with clustering. We introduce a novel semi-supervised ensemble classifier named Random Forest of Tensors (RFoT) that is based on generating a forest of tensors in parallel, which share the same first dimension, and randomly selecting the remainder of the dimensions and entries of each tensor from the features set.

That is, for a given d dimensional tensor $\mathcal{X}$ shaped $n_1 \ x \ n_2 \ x \ \cdots \ x \ n_d$, where the first dimension represents each of the n_1 instances in the dataset, the Canonical Polyadic (CP) tensor decomposition can be written as $\mathcal{X} \approx \mathcal{M} = [\![\lambda \ ; \ \mathbf{A}^{(1)}, \mathbf{A}^{(2)}, \cdots, \mathbf{A}^{(d)}]\!] \equiv \sum_{r=1}^{R} \lambda_r \mathbf{a}_r^{(1)} \circ \mathbf{a}_r^{(2)} \circ \cdots \circ \mathbf{a}_r^{(d)}$, where $\mathcal{M}$ is the low-rank approximation of $\mathcal{X}$ and $\mathbf{A}^{(d)} = [\mathbf{a}_1^{(d)}, \mathbf{a}_2^{(d)}, \ldots, \mathbf{a}_R^{(d)}]$ is the set of R latent factor vectors for dimension d. Using CP-ALS tensor factorization [3, 2, 4], we can group together the samples of one class in each of the R factor vectors $\mathbf{a}_r^{(1)} \in \mathbb{R}^{1 \,\mathrm{x}\, \mathrm{n}_1}$ for the first dimension. Because patterns discovered by CP-ALS can differ for different configurations of tensor dimensions and values over a dataset, we employ the *wisdom of crowds* philosophy to make use of the decisions made by the majority of the randomly generated tensors with varying dimensions and entries. Each tensor configuration decomposed with a randomly selected rank will obtain unique perspectives allowing for the discovery of different latent information. Therefore, if a specific tensor configuration yields poor groupings among classes, the effect would be negligible compared to tensors where meaningful arrangements among classes are discovered. We can then apply the Gaussian Mixture Model clustering algorithm to

*meren1@umbc.edu

†nicholas@umbc.edu

capture the patterns at each of the latent factors for the first dimension $\mathbf{a}_r^{(i,1)}$ within each R_i component, for each of the i tensor configurations $\mathcal{X}^{(i)}$. We employ the cluster *purity score* threshold based on only the known samples to remove the noisier components. Each r component of the ith tensor votes on the sample classes over $c_{i,r}$ clusters using the handful of known instances in a semi-supervised way. Final class prediction can then be obtained by performing a majority vote on each sample.

We show that this methodology yields precise classification results using a small number of known samples to label the rest, making RFoT a novel solution for problems which little in the way of labelled data. We first apply this method on the EMBER-2018 [1] dataset to classify malware and benign-ware using PE header information in the executables. Our preliminary results gave examples where the labeled data do not provide useful information about the patterns detected, and we cannot make classification decisions in those cases. However, in cases where we can make decisions, we achieved F1 scores above 0.97. Finally, we note that RFoT can generalize to other datasets including multi-class classification problems.

References

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

[2] 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.

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

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