Making classifiers more trustworthy using trust scores

Habib Daneshpajouh[1], Kasra Jamshidi[1], Sean La[1,2], Joseph Lucero[3], Matthew Nguyen[1,4]

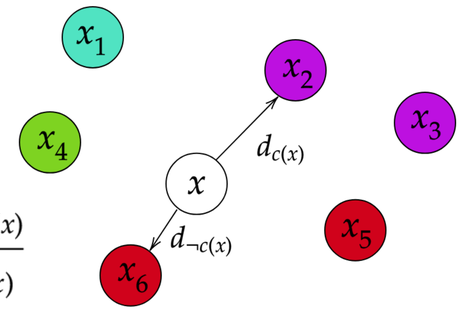
Simon Fraser University

School of Computing Science[1], Departments of Mathematics[2], Physics[3], Molecular Biology & Biochemistry[4]

Goal

Improve the trustworthiness of various classification algorithms using trust scores.

Background



Trust score:

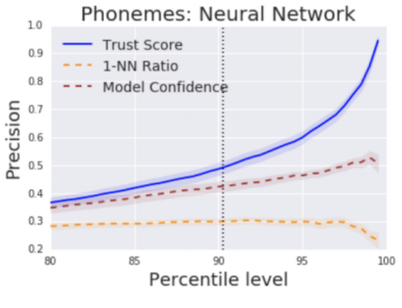
$$t(x, c(x)) = \frac{d_{\neg c(x)}}{d_{c(x)}}$$

The trust score measures the ***trustworthiness*** of a classifier's decision *c(x)* by comparing the distance of the classification with the distance to the nearest neighbor not of that class.

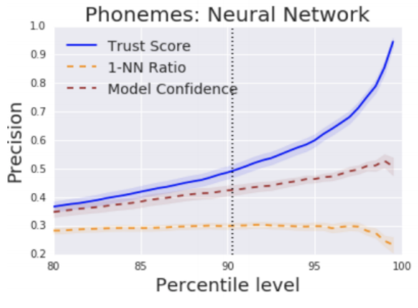
Higher trust scores are better!

Trust scores have been shown to predict correct and incorrect classifications well...

Detect Suspicious



Detect Trustworthy



...and have nice theoretical properties [1].

Theorem 1 (Algorithm 1 guarantees). *Let $0 < \delta < 1$ and suppose that f is continuous and has compact support $\mathcal{X} \subseteq \mathbb{R}^D$ and satisfies Assumption 1. There exists constants $C_l, C_u, C > 0$ depending on f and δ such that the following holds with probability at least $1 - \delta$. Suppose that k satisfies $C_l \cdot \log n \leq k \leq C_u \cdot (\log n)^{D(2\beta+D)} \cdot n^{2\beta/(2\beta+D)}$. Then we have*

$$d_H(H_\alpha(f), \widehat{H}_\alpha(f)) \leq C \cdot \left(n^{-1/2D} + \log(n)^{1/2\beta} \cdot k^{-1/2\beta}\right).$$

Methods

Approach 1:

Use trust scores as weights in a modified 1-Nearest Neighbor algorithm.

1. Compute trust scores for each class,

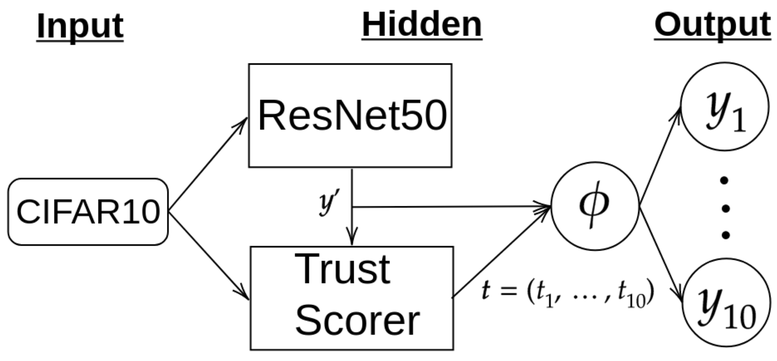
$$t = [t_1 \; t_2 \; \dots \; t_{10}]$$

2. Choose class that has greatest trust score.

$$c(x) = \arg\max_i [t_1 \; t_2 \; \dots \; t_{10}][i]$$

Approach 2:

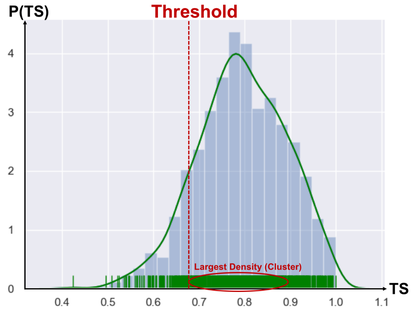
Use trust scores to weight the output of black box neural networks.



$$\phi(y, t) = y' \circ softmax(t)$$

Approach 3:

Use distribution of trust scores to determine a threshold.

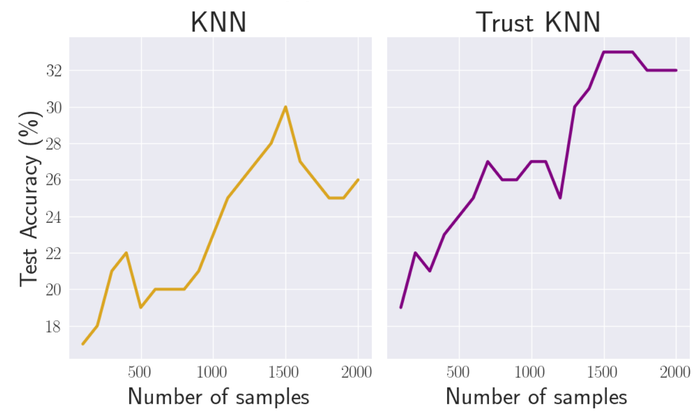


Distribution of trust scores of correctly classified examples

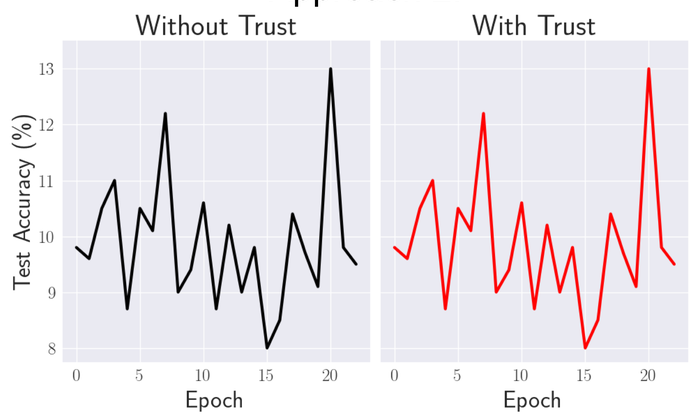
1. Find the largest cluster of trust scores,
2. Threshold is the left-most member of this cluster.
3. Trust scores below the threshold are not acceptable.

Results

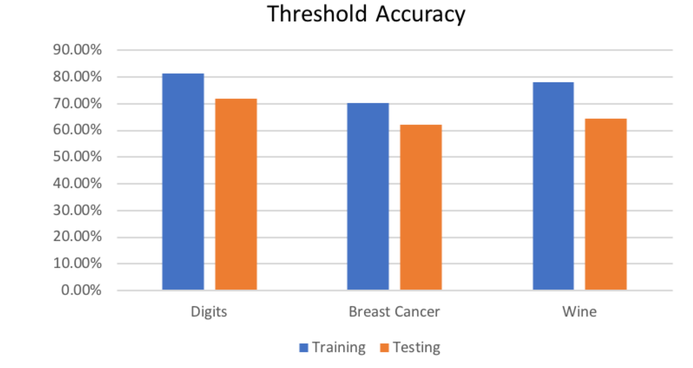
Approach 1:



Approach 2:



Approach 3:



Reference

1. Heinrich Jiang, Been Kim, Melody Y. Guan, and Maya Gupta. *To trust or not to trust a classifier. NeurIPS 2018.* arXiv:1805.11783