Inferring Class Label Distribution of Training Data from Classifiers: An Accuracy-Augmented Meta-Classifier Attack

Raksha Ramakrishna, György Dán, KTH Royal Institute of Technology, Stockholm, Sweden

(rakshar, gyuri}@kth.se)

Class-Label Distribution Inference

Class-label distribution of training data:

\[ p = \frac{N_C}{\sum_{c=1}^{C} N_C} \]

- Sensitive information in banking, manufacturing, health
- Question: Can adversary infer class label distribution based on trained model?

Ingredients for class-label distribution inference:

- Target classifier \( f_c(x; \theta) \) trained on class-label distribution \( p_z \)
- For inference: parameter \( \theta \), accuracy \( a \) over auxiliary dataset \( D_{aux} \) with \( N_{aux} \) samples from each class used
- Metric of accuracy

\[ D_{KL}(p||\hat{p}) = \sum_{c=1}^{C} p_z \log \left( \frac{p_z}{\hat{p}_c} \right) \]

Meta-Classifier Training

- Meta classifier \( \hat{p} = g_{MC}(\theta; a; \omega_{MC}) \) parameters learned via shadow-training
- Shadow-training datasets \( (SD_k, p_k) \) to train shadow classifiers \( f_k \)
- Use parameters \( \theta_k \) and accuracy \( a_k \) to learn meta-classifier parameters \( \omega_{MC} \)

Numerical Results

- Binary classification: Class-label distribution is Bernoulli (\( p \))
- Accurate estimates for most values of \( p \) especially very imbalanced datasets
- Shadow-classifiers trained using class-label distributions with different step sizes (\( \Delta p \))
- Large improvements over baseline (Ganju et.al)

Potential Countermeasure

- Random oversampling of minority class: address class-imbalance
- Makes class-label distribution uniform
- Meta-classifier can still estimate original distribution!
- Further training on oversampled datasets improves performance

Future Work

- Develop new meta-classifiers specific to other target architectures like CNNs
- Mitigation measures
- Extension to Federated Learning

References