ISS-SP-043

識別器差分による識別器クローン手法

安藤 申将[1]　伊藤 千紘[1]　鏡川 悠介[1]　皆川 哲範[1]　酒造 正樹[2]　前田英作[2]
東京電機大学 1 情報環境学部　2 システムデザイン工学部

1. はじめに

近年、機械学習を利用して画像等の認識を行うサービス(MLaaS) が広く公開されている。サービス内部の識別器はブラックボックスだが、その入出力をサンプリングし学習することで識別器のクローンが作成可能である。これはモデル抽出攻撃と呼ばれる[1]。本稿においてはモデル抽出攻撃の危険性を検証するため、効率的な識別器のクローンアルゴリズムを提案する。

2. 提案手法

効率的なクローン処理において、ターゲットとなる識別器をより少ないサンプリング回数でクローンするためには、識別器に入力する有効性が高い特徴量を探ることが重要である。本稿で提案するアルゴリズムを以下に説明する。

サンプリング回数 n が与えられたとき、i = 1, 2, 3,⋯, n まで繰り返す。i - 1 回目までのサンプリングによって、ターゲット識別器に入力された特徴量と出力ラベルのデータセットが存在している。このデータセットを識別器 A と識別器 B に独立に学習させる。ターゲット識別器に入力する特徴量の候補が多数存在する。候補のラベルを学習済みの識別器 A と識別器 B に独立に予測させる。多数の特徴量の中から入力する特徴量を選択するためには、特徴量の有効度を算出する必要がある。識別器 A と識別器 B の予測結果の差分が大きいほど、入力特徴量の候補が識別境界に近く、入力として有効である予想される。したがって、差分が最も大きい特徴量を、 i 回目の入力特徴量として選択する。

3. 実験

データセットは iris, wine, breast_cancer, digits の 4 種類を用いた。ランダムにサンプリングした場合と提案アルゴリズムによるサンプリングの正解率(Accuracy)を比較する。ランダムサンプリングとは、サンプリング回数 n が与えられたとき入力特徴量候補の集合の中からランダムに n 個の特徴量を選択する手法である。各データセットに対して 10 回交差検定を行った結果を表 1 に示す。ランダムサンプリングより提案アルゴリズムを用いたサンプリングの方が高い正解率を示している。

4. 考察

実験結果から、低次元（iris:4）から高次元（digits:64）データセットに対する識別器のクローンにおいても、提案手法は、一般によく用いられるランダムサンプリングに比べて高い正解率を示していた。本手法が有効である理由として、ターゲット識別器がもつ特徴空間付近の特徴量を重点的にサンプリングすることと、入力特徴量同士の距離が離れていることの 2 つが重要であった。

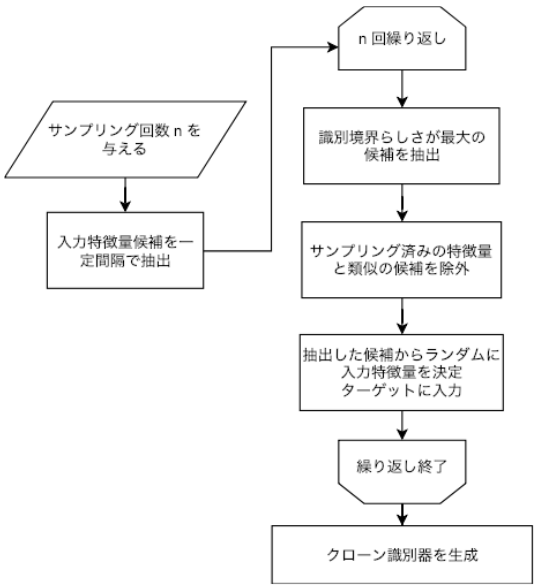

図 1. 提案アルゴリズムの実行フロー

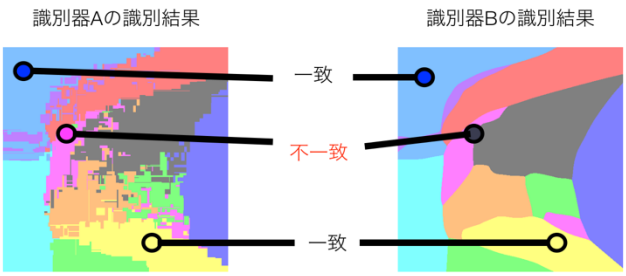

図 2. 特徴空間同士の差分による識別境界らしさの算出

表 1. 実験結果

データセット	次元	ランダム	提案手法
iris	4	0.912	0.947
wine	13	0.965	0.997
breast_cancer	30	0.968	0.992
digits	64	0.404	0.461

参考文献

[1] F. Tramer, *et al.*, “Stealing Machine Learning Models via Prediction”, in Proc. 25th USENIX Security Symposium, pp. 601-618, 2016.