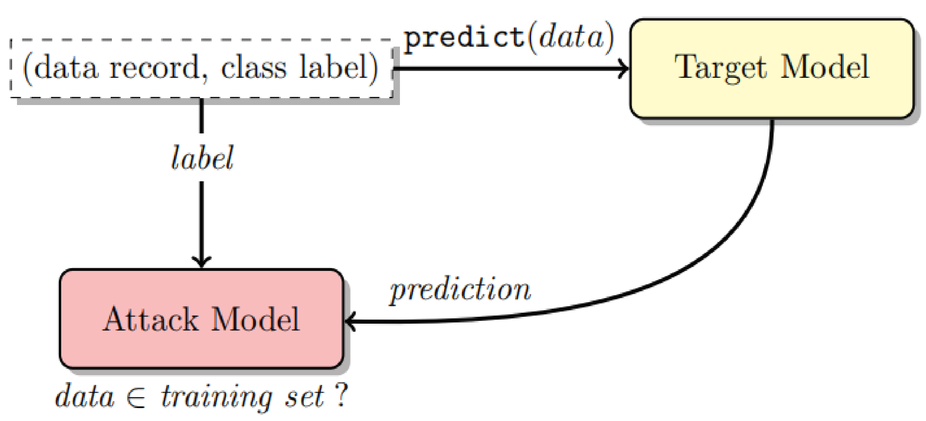
Machine Learning Security: Inference/Membership Attack

LÃ NGỌC ÁNH – 20521065, TRẦN ĐẠI DƯƠNG - 20521226, TRƯƠNG ĐÌNH TRỌNG THANH - 20520766
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN - ĐHQG TP.HCM

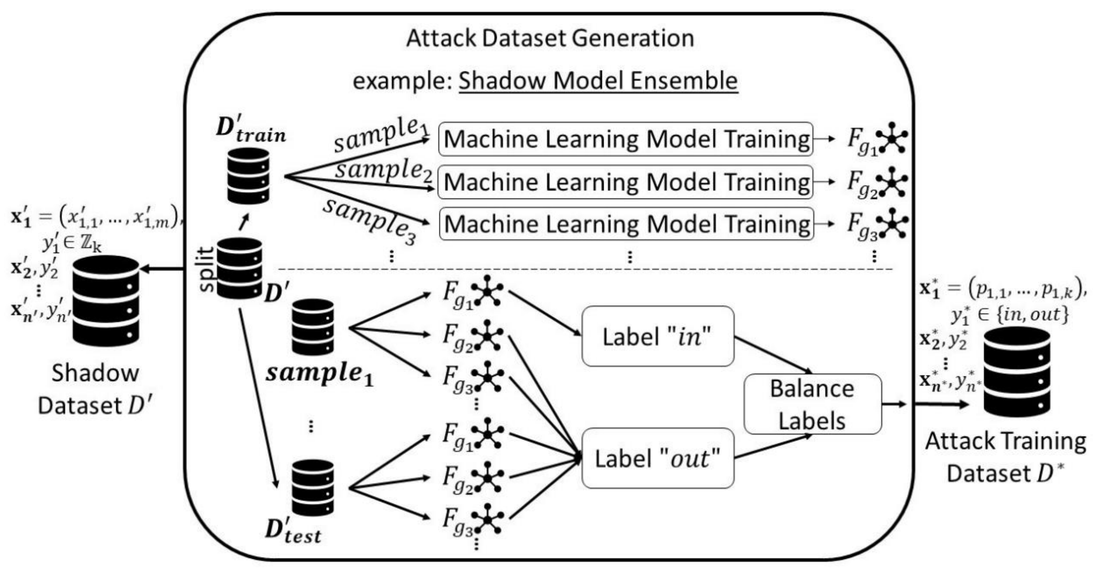
Giới thiệu

Dịch vụ học máy đã chứng kiến một sự bùng nổ quan tâm đến sự phát triển đến từ nền tảng đám mây dịch vụ. Amazon, Microsoft , IBM và Google tất cả đã đưa ra các hình thức máy học máy học như một dịch vụ. Các dịch vụ này cho phép các công ty tận dụng công nghệ máy học hiệu quả và trí tuệ nhân tạo mà không yêu cầu kiến thức chuyên môn. Các nền tảng máy học như một dịch vụ cho phép người dùng tải lên dữ liệu của họ, chạy các phân tích dữ liệu khác nhau hoặc các chuyên gia xây dựng mô hình và triển khai các mô hình được đào tạo cho các dịch vụ của riêng họ.

Với bối cảnh này, mối quan tâm mới đã được trao cho các lỗ hổng tiềm ẩn của các dịch vụ học máy như vậy. Một trong những lỗ hổng đó là Membership inference.



Hình.1 *mô hình dự tính*



Hình.2 *Minh họa quá trình hình thành MI*

Phương pháp

Mục tiêu của kẻ tấn công là xây dựng một mô hình tấn công có thể nhận ra những khác biệt như vậy trong hành vi của mô hình mục tiêu và sử dụng chúng để phân biệt các thành viên với những thành viên không phải là thành viên của tập dữ liệu đào tạo của mô hình mục tiêu chỉ dựa trên đầu ra của mô hình mục tiêu.

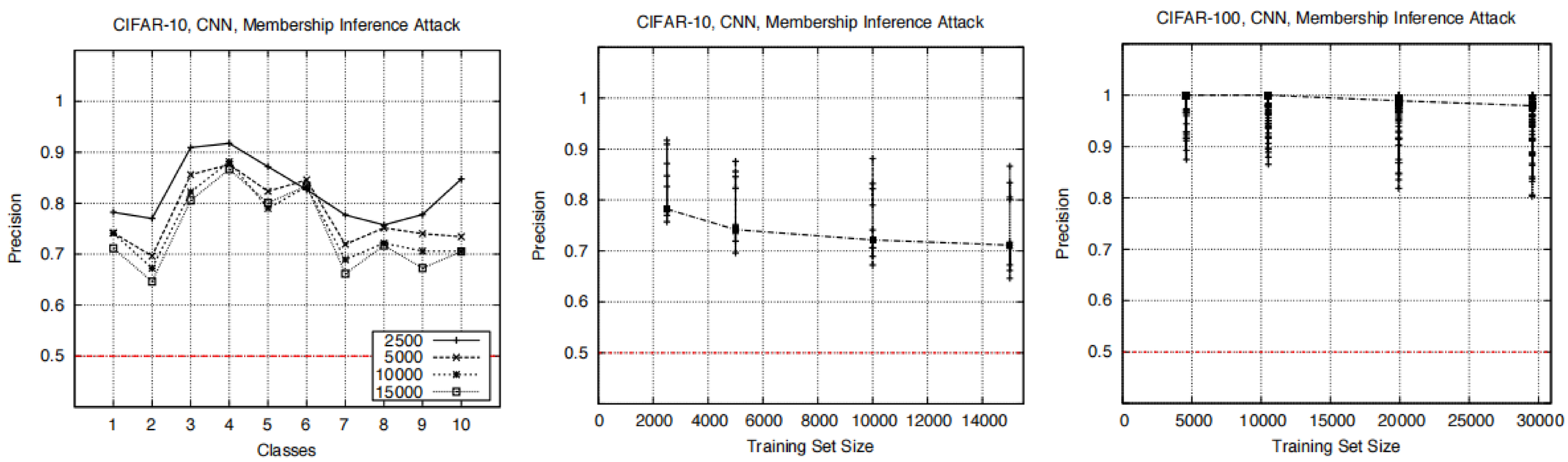
Kẻ tấn công truy vấn mô hình đích bằng một bản ghi dữ liệu và thu được dự đoán của mô hình trên bản ghi đó. Dự đoán là một vectơ của xác suất, một cho mỗi lớp, rằng bản ghi thuộc về một lớp nhất định.

Vectơ dự đoán này, cùng với nhãn của bản ghi đích, là được chuyển đến mô hình tấn công, điều này cho biết liệu bản ghi có nằm trong hoặc ra khỏi tập dữ liệu đào tạo của mô hình mục tiêu.

Kết quả

Sơ đồ miêu tả độ chính xác của Membership inference attack chống lại tập dataset CIFAR 10 và CIFAR 100 được huấn luyện dựa trên model mạng nơ rông tích chập.

Đối với từng kích thước của dữ liệu huyến luyện khác nhau từ 2500,5000, 10000,15000. Các độ chính xác trung bình (từ kích thước tập dữ liệu nhỏ nhất đến lớn nhất) là 0,78, 0,74, 0,72, 0,71 đối với CIFAR-10 và 1, 1, 0,98, 0,97 đối với CIFAR-100.



Hình.3 *Độ chính xác của cuộc tấn công suy luận thành viên chống lại các mạng thần kinh được đào tạo trên bộ dữ liệu CIFAR*