機械学習によるSQLインジェクションの有効な特徴量の探索
Searching for Effective Features of SQL Injection Using Machine Learning

研究駆動コース　29R　林　美里

1. 研究背景・目的

英語の文章は単語と文法構造で意味を表すが、

SQL文は記号を使って条件を指定する

英文	My name is Hayashi and I like electronics.
SQL文	SELECT * FROM users WHERE name = 'Hayashi' AND hobby = 'electronics';

➡ SQLインジェクションの検出において、

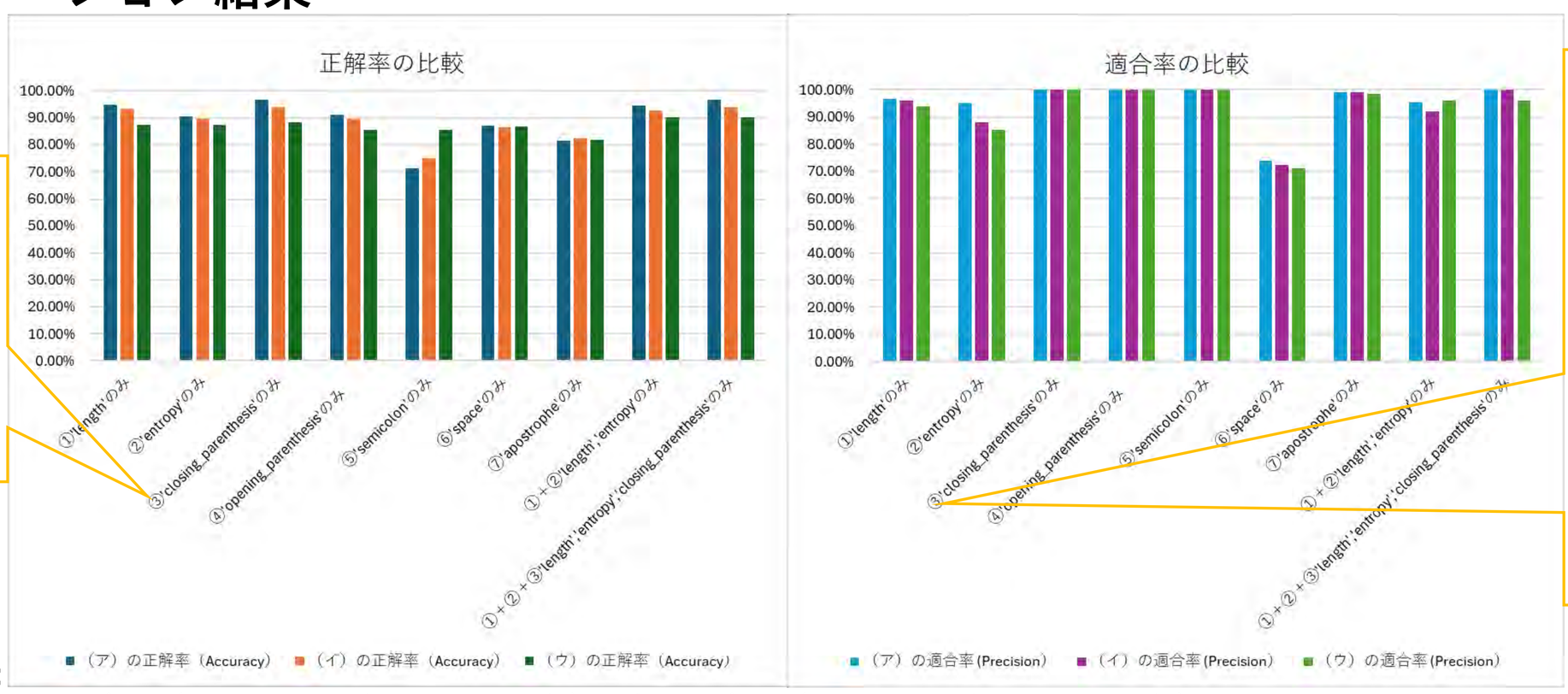
クエリ内の記号（特殊文字）の有無は重要なポイントになり得る

- SQLインジェクション攻撃(SQLi)により、個人情報の搾取や改ざんの被害が続いている[1][2]。
- これまでの研究では、HTTPクエリ文字列の**長さ**、**エントロピー**、**閉じ括弧の有無**を特徴量として使用[3]。ここで、**エントロピー** $H(X) = -\sum_{i=1}^{n} P_i \log_2 P_i$
  （P_i:HTTPクエリ文字列中の文字の出現確率）
- 攻撃検出の精度向上には「適切な特徴量の選定」が不可欠。適切な特徴量を探索し、攻撃と通常のクエリを高精度に分類することが重要。本研究では 「どの特徴量がSQLi検出に有効か？」を検証。

2. 研究方法

- データセット: HttpParamsDataset(異常ラベル付きHTTPクエリ文字列を含む)[3]と園田先生提供のデータセット[2]
- 仮定:
  - SQLiのHTTPクエリ文字列は、特定のSQL文法に基づき、通常のクエリと異なるエントロピーを持つ。
  - SQLi特有の記号の出現頻度を考慮[2]。
- 特徴量エンジニアリング:
  - エントロピーを算出し、数値化。
  - 参考文献[3]のTrainデータ(sqli:7,235行)中での閉じ括弧の出現割合と、参考文献[2]の園田先生提供のTrainデータ(sqli:625行)中での閉じ括弧の出現割合を調査 → それぞれ90.91%、45.92%であった。
- 検出器開発（DecisionTreeClassifier(決定木)を使用）:
  - length, entropy, closing_parenthesisなどの組み合わせでSQLiの検出器を作成。精度評価を実施。

3. シミュレーション結果



・①から⑦の組み合わせで正解率（Accuracy）が高いものを探した。
③閉じ括弧のみでも高い正解率が得られた。

・Web アプリケーションファイアウォール（WAF）では誤検知 (False Positive) が問題となる。誤検出しない確率である適合率（Precision）を見てみると、③閉じ括弧のみでも高い適合率が得られた。

使用データセット:
(ア) 参考文献[3]のTrainデータ(norm:12,870行, sqli:7,235行)及びTestデータ(norm:6,434行, sqli:3,617行)。
(イ) 参考文献[3]のTrainデータの先頭からnorm: 1,112行を抽出したTrainデータ(norm:1,112行)と参考文献[2]の園田先生提供のTrainデータ(sqli:625行)及び参考文献[3]のTestデータ(norm:6,434行, sqli:3,617行)。参考文献[2]と参考文献[3]の両方からnormとsqliの比率を変えずにTrainデータを作成。
(ウ) 参考文献[3]のTrainデータの先頭からnorm: 1,112行を抽出したTrainデータ(norm:1,112行)と参考文献[2]の園田先生提供のTrainデータ(sqli:625行)及び参考文献[3]のTestデータの先頭からnorm: 556行とsqli:312行を抽出したTestデータ(norm:556行, sqli:312行)。参考文献[3]のTrainデータ(sqli:7,235行)と参考文献[2]のTrainデータ(sqli:625行)のsqliの比率に合わせて参考文献[3]のTestデータのデータ数を削減。

- 結果:
  - closing_parenthesis のみでも高い正解率 (Accuracy) を達成。
  - 適合率 (Precision) においても closing_parenthesis のみで高い値を確認。

4. 結論

- 機械学習モデルの性能は入力データ（特徴量）に大きく依存。SQLiの特徴を考慮した新たな特徴量を提案・評価。
- Webアプリケーションのセキュリティ強化に貢献する可能性を示した。

参考文献

[1] IPA 脆弱性対策情報データベース「CWE-89 SQLインジェクション」
[2] 園田道夫, 中央大学博士論文「潜在的要因を考慮したSQLインジェクション攻撃検知システムの開発」
[3] Chiheb Chebbi 他, 「セキュリティエンジニアのための機械学習」, オライリージャパン