4W-06

攻撃分類を行う侵入検知システムにおける能動学習の性能評価

赤坂省伍† 梅澤猛‡ 大澤範高‡
千葉大学工学部† 千葉大学大学院融合科学研究科‡

1. はじめに

侵入検知システムにおいて、高精度に攻撃を分類するモデルを機械学習によって構築することができれば、ネットワークセキュリティを大きく向上させることができる。そのためには、ラベリング(正答ラベルを付加)された学習データが大量に必要となるが、一般にラベリングは手動で行われるため、作業コストが高くなるという問題がある。

また、攻撃手法は多種多様であり、攻撃手法ごとに異なる対策を取る必要があることから、攻撃か否かだけではなく、攻撃手法を判別することが重要である。

そこで本研究では、能動学習によってデータを選択的に利用することで学習効率の向上を図り、攻撃の分類を行う機械学習型侵入検知システムにおいて、少量のデータで高精度モデルを構築可能とすることでラベリングコストの低減を図る。

2. 関連研究

Revathi らは Random Forest、J48、Support Vector Machine (SVM)、CART、Naïve Bayes を用いて、攻撃分類を行う侵入検知の検証を行っている[1]。検証には 25,192 件の学習データを使用しており、同様に侵入検知システムを構築することを考えると、同程度の件数を手動でラベリングする必要があるためラベリングコストが高いという問題がある。

Chairi らは SVM を用いた侵入検知に能動学習を適用し、少量の学習データで高精度な分類が可能かを検証している [2]。検証は攻撃か否かの検知で行われており、攻撃手法の判別を行っていないため、攻撃分類を行う際の検証が必要である。

3. 実験

能動学習の有無による分類精度への影響を調査するため、Margin Sampling(手法 M)、Least Confident(手法 L)、Entropy(手法 E)の 3 つのデータ選択手法に、比較用として無作為なデータ選択を加えた計 4 つのモデルによる比較実験を行った。学習データには NSL-KDD の Train+20% subset (25,192 件)を、評価データには NSL-KDD の Test set (22,544 件)を用いた。これらのデータは 41 次元の特徴量を持ち、40 のクラスがラベルとして付加されている。本研究では、あらかじめラベルを 5 つのクラス(Normal、DoS、R2L、U2R、Probe)に大別して調査を行った。学習アルゴリズムには SVM を使用した。

分類モデルの更新手順を図 1 に示す。まず、能動学習の有無に関係なく、5 種類のクラスから 1 件ずつ計 5 件を無作為に抽出して最初のラベル付きデータとして分類モデルを構築する。構築したモデルとデータ選択手法をラベル無しデータに適用して追加データを抽出し、ラベルを付加してラベル付きデータに追加する。以上の手順をラベル無しデータが無くなるまで繰り返す。

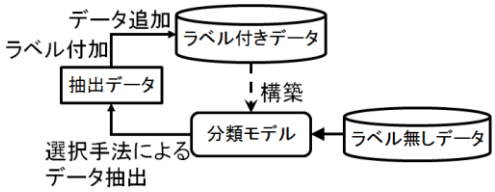

図 1：学習モデル更新の手順

分類モデルの精度評価は、モデル構築のたびに評価データを用いて行った。初期モデルの性能により結果が偏ることを避けるため、最初のラベル付きデータを 1,000 通り作成し、学習と評価を行って平均を求めた。

4. 結果と考察

各手法を用いてモデルの更新を繰り返した結果を図 2 に示す。また、手法ごとの標準偏差付きの結果を図 3 ～ 図 6 に示す。指標としては、再現率と適合率の調和平均によって総合的な精度を示す F 尺度を用いた。縦軸と横軸は、それぞれ F 尺度と使用データ数を示している。能動学習無しのモデルにおいて全データを使用した時の F 尺度の値の 90.0%を閾値とし、閾値以上の分類精度を得るために必要なデータ数を比較し、一定以上の分類精度を得る速さを評価した。図 3 ～ 図 6 中のバーは標準偏差を表す。

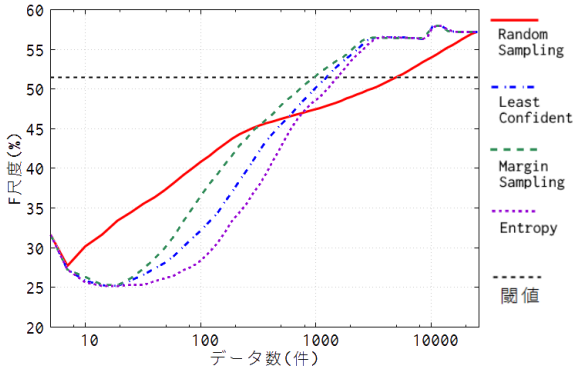

図 2：能動学習の有無による F 尺度のマクロ平均

An Evaluation on Active Learning for Attack Classification in Intrusion Detection System
†Shogo Akasaka, Faculty of Engineering, Chiba University
‡Takeshi Umezawa , Noritaka Osawa, Graduate School of Advanced Integration Science, Chiba University

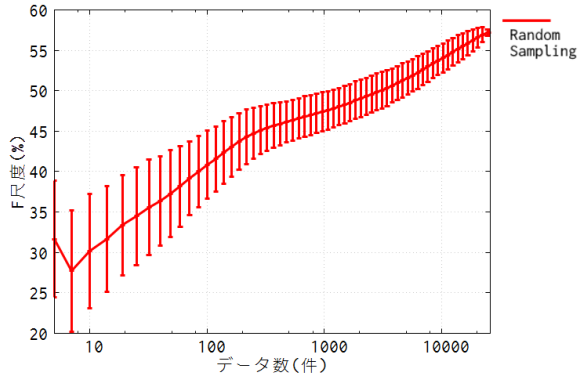

図 3：能動学習無しの F 尺度の標準偏差

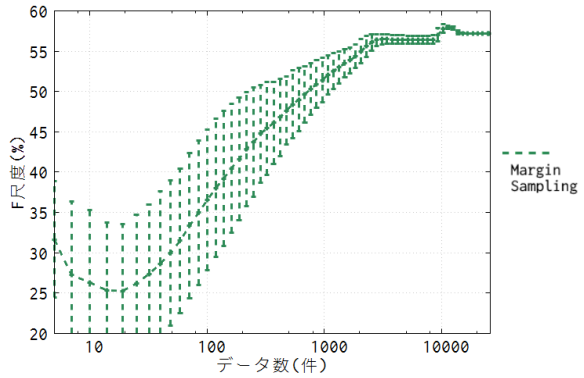

図 4：手法 M での F 尺度の標準偏差

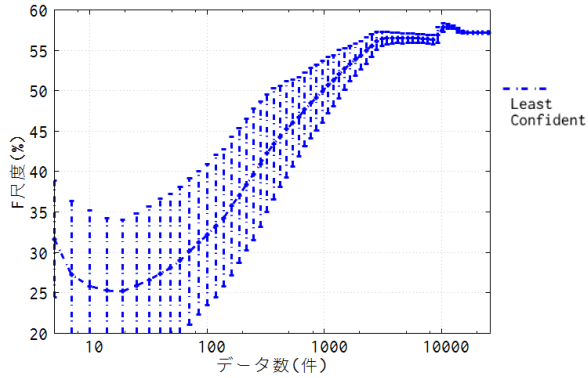

図 5：手法 L での F 尺度の標準偏差

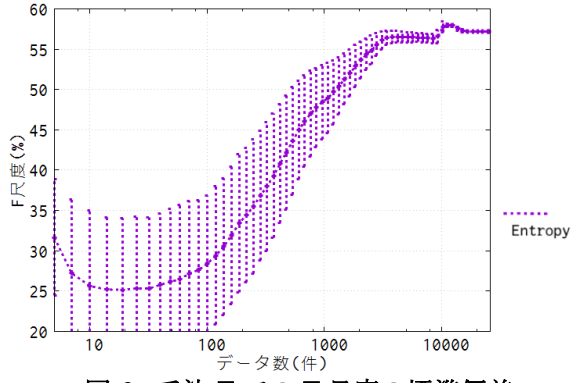

図 6：手法 E での F 尺度の標準偏差

図 2 より、学習データ数は手法 M で 325 件、手法 E で 767 件以上の時、能動学習有の方が高い精度を示した。また、能動学習を適用した場合の F 尺度の値はデータ数が手法 M で 1,077 件、手法 E で 1664 件の時に閾値を超えた。一方、能動学習無しの場合には、データ数が 5,145 件の時に閾値を超えた。したがって、学習データ数が 325 件以上の時は能動学習を利用する方がよく、能動学習を適用することで閾値に達するのに必要な学習データ数を無作為な手法に比べて 4,068 件削減できることが示された。また、分類精度の収束の早さを標準偏差で比較すると、学習データ数が 1,203 件以下では能動学習無しの方が小さい標準偏差を示しているが、1,203 件以降は能動学習有の方が小さくなっている。さらに、変動係数が 1%未満となる時に標準偏差が収束したとし、データ数を比較すると能動学習無しの場合はデータ数が 24,690 件で標準偏差が収束しているが、能動学習有の場合、手法 M で 4,210 件の時に収束している。したがって、能動学習を適用することで標準偏差が収束するのに必要な学習データ数が減少しており、能動学習の優位性が示された。しかし、学習データ数が 325 件以下では能動学習無しの方が高い値を示しており、データ数が 1,077 件以下の箇所では能動学習無しの方が変動係数は低い値であるため、データ数が少ないときの改善が必要である。

5. まとめ

機械学習型侵入検知システム構築の問題点である学習データのラベリングコスト削減を目指し、能動学習による分類モデルを評価することで学習データ数削減への有効性を確認した。学習データ数で 325 件以上のとき能動学習有の方が高い精度を示し、能動学習により全学習データ使用時の F 尺度の 90.0%に設定した閾値を超えるのに必要な学習データ数を最大 4,068 件削減可能なことを示した。また、標準偏差で比較すると、収束するのに必要な学習データ数が減少しており、優位であることを示した。今後の課題としては、学習データ数が少ない場合に能動学習無しの方が高い値を示す問題を改善する手法の検討が挙げられる。

参考文献

[1] Revathi, S. and Malathi, A., “A detailed analysis on NSL-KDD dataset using various machine learning techniques for intrusion detection”, International Journal of Engineering Research and Technology, Vol. 2, Issue 12, pp. 1848-1853 (2013).

[2] Chairi, I., Alaoui, S., and Lyhyaoui, A., "Sample Selection Based Active Learning for Imbalanced Data", 2014 Tenth International Conference on Signal-Image Technology and Internet-Based Systems, pp. 645-651 (2014).