侵入検知モデル構築における能動学習と無作為選択を用いたラベリングコスト削減手法の検討

赤坂 省伍[†1]　梅澤 猛[†2]　大澤 範高[†2]

概要：機械学習を用いた侵入検知システム構築には，大量の教師データが必要となるため個々のデータにラベルを付ける手作業に高いコストがかかる．筆者らは，学習効率の高いデータを選択的に利用する手法である能動学習を適用してラベリングコストの削減を試みた．しかし，学習初期の教師データ数が少ない状態においては，能動学習による効果がみられず，無作為にデータを選択した場合より精度が低くなるという問題が確認された．そこで本研究では，極少量の教師データを用いて初期の分類モデルを構築し，学習済みデータが少ない時期には無作為にデータ選択を行い，途中から能動学習によるデータ選択に切り替える手法を用いて学習の進行状況に依らずラベリングコストを削減可能か検討する．常に無作為にデータを選択した場合を比較対象として分類精度を評価し，同程度の精度を得るのに必要な教師データの数を調査した．

キーワード：侵入検知システム，機械学習，能動学習，NSL-KDD，UNSW-NB15

1. はじめに

ネットワークからの攻撃を検知する侵入検知システムを機械学習により構築するには，大量の教師データが必要となる．教師データは，過去に観測されたデータに対し人手で正答ラベルを付与（ラベリング）する必要があるため，大量の教師データを準備するには高いコストがかかる．

この問題を解決するために，筆者らは能動学習を用いたデータ選択による分類モデルと，無作為なデータ選択による分類モデルの比較評価を行い，能動学習の適用でラベリングコストが削減できることを確認した [1]．このときの評価においては，極少量の教師データから各データ選択方法を用いて分類モデルの構築を行ったが，能動学習による分類モデルが常に高い分類精度を示すわけではなく，教師データ数が少ない状態においては，無作為なデータ選択による分類モデルの方が高い精度を示した．この教師データ数が少ない状態において能動学習による分類精度が低い問題は，検証で用いた能動学習アルゴリズムが学習効率の高いデータとして分離境界に近いものを選択するものであったため，極少量の教師データで構築した分類モデルの分離境界が実際の分離境界と著しく離れており，教師データ数が少ない状態において能動学習が上手く働かなかったためだと考えられる．したがって，教師データが少ない状態においては，無作為な選択により教師データを一定まで増やすことで，実際のものに近い分離境界を推定することができ，能動学習の効果が発揮できるようになると考えられる．

そこで，教師データが少ない状態ではデータ選択を無作為に行い，教師データ数が一定以上となったとき能動学習に切り替えることで，教師データ数が少ない状態において能動学習による分類精度が低い問題の改善を行い，ラベリングコストの削減を図る．本研究では，データ選択方法を切り替える基準を複数用意し，分類精度の推移を比較した．

2. 関連研究

Bhavsar らは Support Vector Machine（SVM）を用いた侵入検知の検証により, 98.6 %と高い正解率を得ている [2]．また，Revathi らは Random Forest（RF）等の機械学習手法を用いた侵入検知の検証により，99.8 %と高い正解率を得ている [3]．

これらの検証には 25,192 件の教師データが使用されており，同程度の正解率の侵入検知システムを構築することを考えると，大量の教師データを用意する必要があるためラベリングコストが高くなるという問題がある．

3. 能動学習

能動学習は，教師データを作成する際に最も効率的に学習を行うことができると考えられるデータを自動的に選択し，選択されたデータのみに人間がラベリングを行うことで，できるだけ少量の教師データで高精度の分類モデルを構築することを目的とする機械学習手法である [4]．本研究では Uncertainty Sampling と呼ばれるフレームワークのデータ選択方法の 1 つである Margin Sampling を用いた．Uncertainty Sampling は，分離境界面の最も近くに存在するデータが最も不確かであるため，そのデータにラベリングすることで効率的に学習を行うことができると考え，データを選択するフレームワークである．

Margin Sampling では，最も不確かなデータとして，ラベリングされていないプールデータから，各クラスに属する確率の内，最も高い確率と 2 番目に高い確率の差が最も小

†1 千葉大学 大学院融合理工学府
Graduate School of Science and Engineering, Chiba University.
†2 千葉大学 大学院工学研究院
Graduate School of Engineering, Chiba University.

© 2018 Information Processing Society of Japan

さいデータを選択する方法である．

4. データセット

侵入検知用ベンチマークデータセットである NSL-KDD [5]を用いて行った検証 [1]では結果がデータセット依存である可能性が否定できなかったため，本稿では UNSW-NB15 [6] [7] [8]においても検証を行う．

4.1 NSL-KDD

NSL-KDD [5]はカナダの University of New Brunswick において開発された侵入検知用ベンチマークデータセットで，40 クラスにラベル付けされおり，本研究では評価のために 4 クラス（Normal，DoS，Exploit，Probe）に大別した．データ数は先行研究 [2] [3]と同様に学習データは 25,192 件，評価データは 22,544 件とした．それぞれのクラスとデータ件数を表 1 に示す．

表 1 : NSL-KDD のクラスとデータ件数

NSL-KDD			
4 クラス	40 クラス	学習データ	評価データ
Normal	Normal	13,449 件	9,711 件
DoS	apache2，back，teardrop，smurf，land，mailbomb，neptune，pod，udpstorm，processtable	9,234 件	7,458 件
Exploit	buffer_overflow，ftp_write，imap，xterm，multihop，guess_passwd，phf，httptunnel，perl, loadmodule，named，rootkit，sqlattack，spy，snmpgetattack，ps，snmpguess，warezclient，xlock，sendmail，xsnoop，worm，warezmaster	220 件	2,954 件
Probe	ipsweep，mscan，nmap, portsweep，saint，satan	2,289 件	2,421 件

4.2 UNSW-NB15

UNSW-NB15 [6] [7] [8]はオーストラリアの University of New South Wales において開発された侵入検知用ベンチマークデータセットで，9 クラスにラベル付けされており，本研究では評価のために 4 クラス（Normal，DoS，Exploit，Probe）に大別した．データ数は先行研究 [7] [8]と同様に学習データは 175,341 件，評価データは 82,332 件とした．それぞれのクラスとデータ件数を表 2 に示す．

表 2 : UNSW-NB15 のクラスとデータ件数

UNSW-NB15			
4 クラス	9 クラス	学習データ	評価データ
Normal	Normal	56,000 件	37,000 件
DoS	DoS	122,641 件	4,089 件
Exploit	Analysis，Worms，Backdoor, Generic，Exploits，Fuzzers，Shellcode	96,586 件	37,747 件
Probe	Reconnaissance	10,491 件	3,496 件

5. 分類モデル構築と教師データ選択の手順

検証で行った分類モデル構築と教師データ選択の手順を図 1 により説明する．

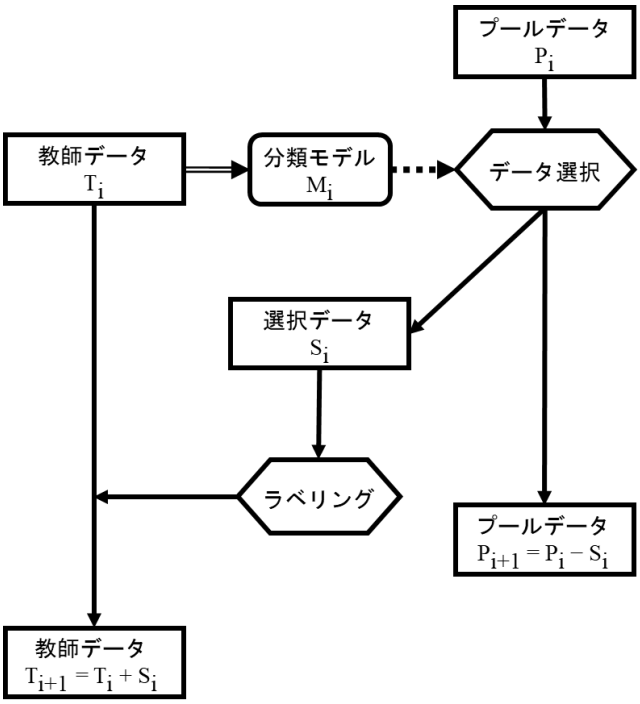

図 1 : 分類モデル構築と教師データ選択の手順

ここでは，i 回目の分類モデル構築と教師データ選択を考える．まず，教師データT_iを用いて分類モデルM_iを構築する．そして分類モデルM_iを用いてラベルの付いていないプールデータP_iを評価し，選択データS_iを選択する．最後に選択データS_iを教師データT_iに加える．これは和集合を+で表すと$T_{i+1} = T_i + S_i$となる．次の教師データにはT_{i+1}を，プールデータにはP_{i+1}を用いる．これは差集合を-で表すと$P_{i+1} = P_i - S_i$となる．評価は分類モデル作成のたびに評価デ

© 2018 Information Processing Society of Japan

ータセットを用いて行う．

最初の教師データ（i = 0）は各クラスから 1 件ずつ，計 4 件を無作為に選択した．また，最初の教師データにより結果が偏ることを避けるため，最初の教師データを 100 通り作成し，それぞれ学習と評価を行い，平均を求める．

6. 評価指標

本研究では，検証の評価指標として再現率と適合率の調和平均である F 尺度を用い，クラスごとに評価を行った．式（ 1 ），式（ 2 ），式（ 3 ）に表 3 を基とした各指標の定義式を示す．N は評価するクラス数，i は評価するクラスである．

表 3：混同行列

		予測結果		
		1	… i …	N
実際	1	E_{11}	… E_{1i} …	E_{1N}
	⋮ i ⋮	E_{i1}	⋮ … E_{ii} … ⋮	E_{iN}
	N	E_{N1}	… E_{Ni} …	E_{NN}

$$再現率_i = \frac{E_{ii}}{\sum_{j=1}^{N} E_{ij}} \qquad 式（1）$$

$$適合率_i = \frac{E_{ii}}{\sum_{j=1}^{N} E_{ji}} \qquad 式（2）$$

$$F\ 尺度_i = \frac{2 \times 再現率_i \times 適合率_i}{再現率_i + 適合率_i} \qquad 式（3）$$

7. 予備実験

あらかじめ SVM と RF を用いて分類し，F 尺度で分類精度を出したところ，分類精度は大きくは変わらず，本研究の実験環境ではモデル構築時間は RF が非常に短かったため，検証は RF を用いて行った． NSL-KDD と UNSW-NB15 の両方において能動学習によるラベリングコスト削減が可能かを調査した．

7.1 NSL-KDD における RF と能動学習の有効性検証

RF を用いた場合でも NSL-KDD において能動学習によるラベリングコスト削減が可能かを調査した．無作為選択（Random Sampling ― 以降，手法 RS）と能動学習（Active Learning ― 以降，手法 AL）を比較する．手法 RS で全学習データ使用時の F 尺度の値の 99.0 %を閾値とし，全クラスにおいて最初に閾値を超えた時の教師データ数を比較し，一定以上の分類精度を得るのに必要な教師データ数を評価した．結果を図 2 に示す．縦軸と横軸はそれぞれ F 尺度と教師データ数を示している．また，閾値を超えるために必要な教師データ数を表 4 に示す．

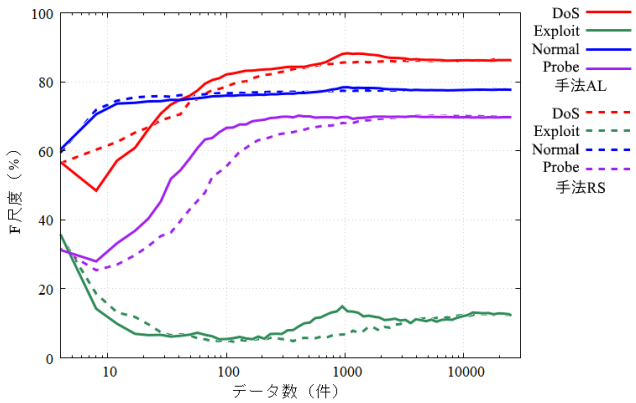

図 2：NSL-KDD – 手法 AL と手法 RS

表 4：NSL-KDD - 閾値を超えるために必要な教師データ数（件）

	DoS	Exploit	Normal	Probe
手法 RS	852	4	186	1,939
手法 AL	692	4	504	234

図 2，表 4 より，手法 AL を用いた場合，教師データ数が 692 件，割合にして 2.75 %の時，全クラスにおいて閾値を超えた．一方，手法 RS の場合，教師データ数が 852 件，割合にして 7.70 %の時，全クラスにおいて閾値を超えた．

したがって，NSL-KDD において RF を用いて分類モデルを構築する際に手法 AL を用いて教師データを選択することで閾値として指定した精度を得るための教師データ件数すなわちラベリングコストを手法 RS と比べて 64.31 %削減できることが示された．

7.2 UNSW-NB15 における RF と能動学習の有効性検証

RF を用いた場合に UNSW-NB15 において能動学習によるラベリングコスト削減が可能かを調査した．無作為選択（手法 RS）と能動学習（手法 AL）を比較する．手法 RS で全学習データ使用時の F 尺度の値の 99.0 %を閾値とし，全クラスにおいて最初に閾値を超えた時の教師データ数を比較し，一定以上の分類精度を得るのに必要な教師データ数を評価した．結果を図 3 に示す．縦軸と横軸はそれぞれ F 尺度と教師データ数を示している．また，閾値を超えた教師データ数を表 5 に示す．

© 2018 Information Processing Society of Japan

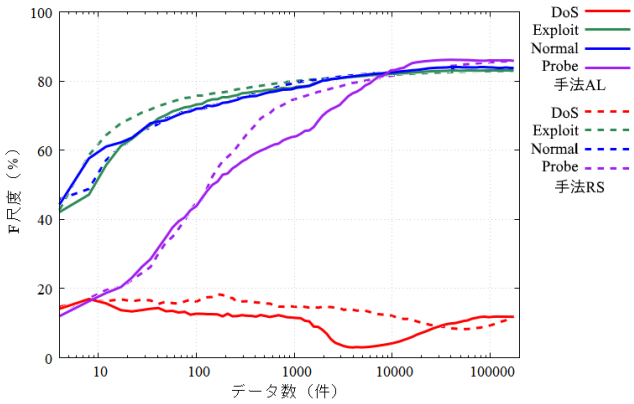

図 3 : UNSW-NB15 – 手法 AL と手法 RS の比較

表 5 : UNSW-NB15 - 閾値を超えるために必要な教師データ数（件）

	DoS	Exploit	Normal	Probe
手法 RS	4	17,769	17,769	72,163
手法 AL	4	11,900	16,075	16,075

図 3，表 5 より，手法 AL を用いた場合，教師データ数が 16,075 件，割合にして 9.17 %の時，全クラスにおいて閾値を超えた．一方，手法 RS の場合，教師データ数が 72,163 件，割合にして 41.16 %の時，全クラスにおいて閾値を超えた．

したがって，UNSW-NB15 において RF により侵入検知システムを構築する際に，手法 AL を用いて教師データを選択することで，指定した閾値を上回る精度を得るためのラベリングコストを手法 RS と比べて 77.72 %削減できることが示された．

しかし，NSL-KDDとUNSW-NB15の両方において，データ数が少ない状態において，手法ALによる分類精度より手法RSによる分類精度の方が高い値を示している（図 2 図 3）．これは，極少量である最初の教師データで構築した分類モデルの分離境界が実際の分離境界と著しく離れており，能動学習アルゴリズムが実際の分離境界から離れたデータを選択してしまい，教師データが少ない状態では手法ALが上手く働かないためだと考えられる．

8. 評価実験

7 節より，NSL-KDD と UNSW-NB15 において RF により分類モデルを構築する際に能動学習を用いることで，ラベリングコストの削減が可能であることが示された．しかし，RF を用いた場合も SVM を用いた場合 [1]と同様に教師データ数が少ない状態において能動学習による分類精度が無作為選択より低い．そこで，教師データが少ない状態では教師データ選択を無作為に行い，教師データ数が一定以上でデータ選択方法を能動学習に切り替えてデータ選択を行う手法（Switch Method ― 以降，手法 SM）の有効性を評価した．

8.1 NSL-KDD における手法 SM の評価と考察

NSL-KDD において手法 SM を評価する．評価はデータ選択方法を切り替える教師データ数が 100 件（Switch Method-NSL-KDD ― 以降，手法 SM-nk1）と 1,000 件（Switch Method-NSL-KDD ― 以降，手法 SM-nk2）の 2 種に対して，無作為選択（手法 RS）との比較を行った．手法 RS で全学習データ使用時の F 尺度の値の 99.0 %を閾値とし，全クラスにおいて最初に閾値を超えた時の教師データ数を比較し，一定以上の分類精度を得るのに必要な教師データ数を評価した．手法 SM-nk1，手法 SM-nk2 の結果をそれぞれ図 4，図 5 に示す．縦軸と横軸はそれぞれ F 尺度と教師データ数を示している．また，手法 SM-nk1，手法 SM-nk2 の閾値を超えるために必要な教師データ数を表 6 に示す．

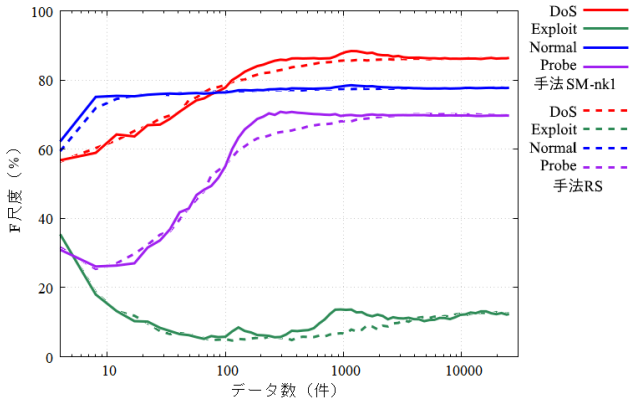

図 4 :手法 SM-nk1 の結果

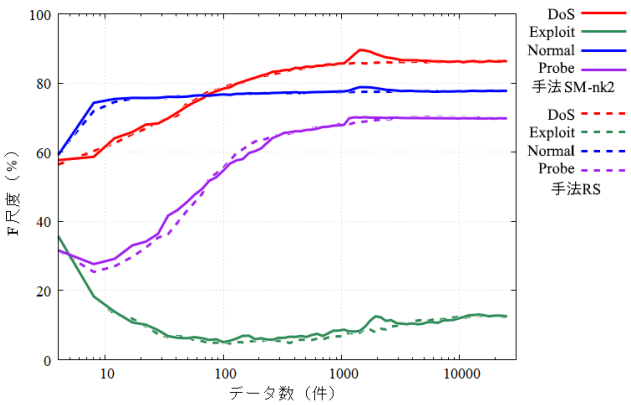

図 5 : 手法 SM-nk2 の結果

© 2018 Information Processing Society of Japan

表 6：閾値を超えるために必要な教師データ数（件）

	DoS	Exploit	Normal	Probe
手法 RS	852	4	186	1,939
手法 SM-nk1	262	4	129	209
手法 SM-nk2	768	4	165	1,162

図 4，図 5，表 6 に示す通り，手法 RS を用いた場合，教師データ数が 1,939 件，割合にして 7.70 %の時，全クラスにおいて閾値を超えた．対して手法 SM-nk1 を用いた場合，教師データ数が 262 件，割合にして 1.04 %の時，全クラスにおいて閾値を超え，削減率は 86.49 %であった．手法 SM-nk2 を用いた場合，教師データ数が 1,162 件，割合にして 4.61 %の時，全クラスにおいて閾値を超え，削減率は 40.07 %であった．

したがって，NSL-KDD において手法 SM を用いた場合でも手法 AL と同様にラベリングコストを削減できることが示された．

図 2，図 4，図 5 より，常に手法 RS による分類精度より手法 SM による分類精度の方が高いもしくは変わらない結果を示した．これらより NSL-KDD において，一定以上まで教師データを無作為に選択することで，教師データ数が少ない状態において能動学習による分類精度が低い問題の解決ができることが示された．

8.2 UNSW-NB15 における手法 SM の評価と考察

UNSW-NB15 において手法 SM を評価する．評価はデータ選択方法を切り替える教師データ数が 1,000 件（Switch Method-UNSW-NB15 ― 以降，手法 SM-un1）と 5,000 件（手法 SM-un2），10,000 件（Switch Method-UNSW-NB15 ― 以降，手法 SM-un3）の 3 種に対して，無作為選択（手法 RS）との比較を行った．手法 RS で全学習データ使用時の F 尺度の値の 99.0 %を閾値とし，全クラスにおいて最初に閾値を超えた時の教師データ数を比較し，一定以上の分類精度を得るのに必要な教師データ数を評価した．手法 SM-un1，手法 SM-un2，手法 SM-un3 の結果をそれぞれ図 6，図 7，図 8 に示す．縦軸と横軸はそれぞれ F 尺度と教師データ数を示している．また，手法 SM-un1，手法 SM-un2，手法 SM-un3 の閾値を超えるために必要な教師データ数を表 7 に示す．

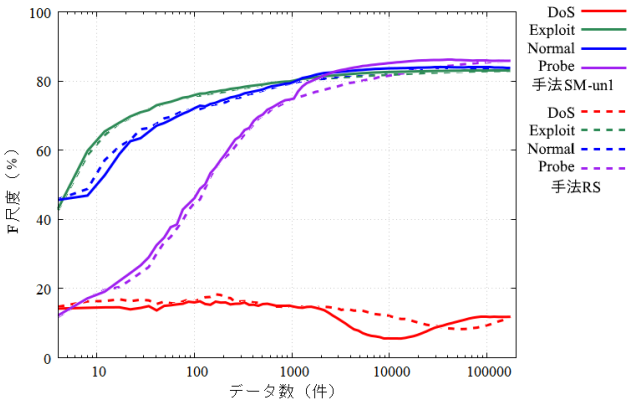

図 6 :手法 SM-un1 の結果

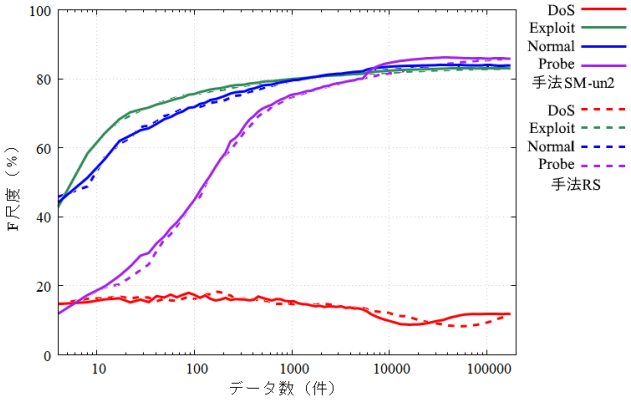

図 7：手法 SM-un2 の結果

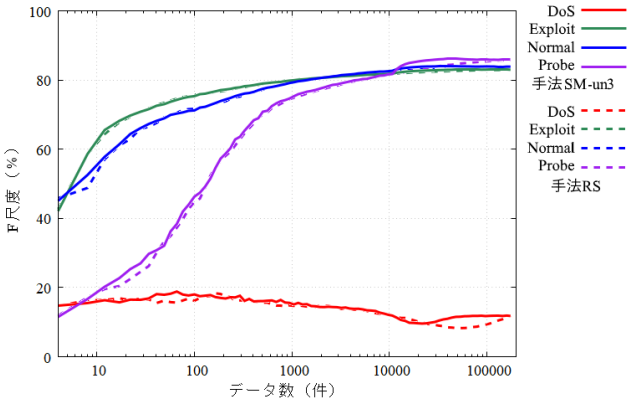

図 8：手法 SM-un3 の結果

表 7 :閾値を超えるために必要な教師データ数（件）

	DoS	Exploit	Normal	Probe
手法 RS	4	17,769	17,769	72,163
手法 SM-un1	4	4,818	4,356	9,736
手法 SM-un2	4	7,965	7,204	13,155
手法 SM-un3	4	13,155	11,900	17,769

図 6，表 7 に示す通り，手法 RS を用いた場合，教師デ

© 2018 Information Processing Society of Japan

ータ数が 72,163 件，割合にして 41.16 %の時，全クラスにおいて閾値を超えた．対して図 7，表 7 に示す通り，手法 SM-un1 を用いた場合，教師データ数が 9,736 件，割合にして 5.55 %の時，全クラスにおいて閾値を超え，削減率は 86.51 %であった．図 8，表 7 に示す通り，手法 SM-un2 を用いた場合，教師データ数が 13,155 件，割合にして 7.50 %の時，全クラスにおいて閾値を超え，削減率は 81.77 %であった．手法 SM-un3 を用いた場合，教師データ数が 17,769 件，割合にして 10.13 %の時，全クラスにおいて閾値を超え，削減率は 75.38 %であった．したがって，UNSW-NB15 において手法 SM を用いた場合でもラベリングコストを削減できることが示された．

また図 6 に示す通り，手法 SM-un1 において 1,000 件から 2,000 件程度まで手法 RS による分類精度より手法 SM の分類精度の方が高いもしくは変わらない結果を示した．また図 7 に示す通り，手法 SM-un2 において 5,000 件から 7,000 件程度まで手法 RS による分類精度より手法 SM の分類精度の方が高いもしくは変わらない結果を示した．これらでは, 教師データを一定まで無作為に選ぶことで手法 RS の方が高い分類精度を示す問題の改善が見られたが，図 6 の 2,000 件から 30,000 件程度までおよび，図 7 の 7,000 件から 30,000 件程度までにおいて，依然として教師データ数が少ない状態では手法 SM による分類精度が低い．図 3，図 8 より，手法 RS による分類精度より手法 SM による分類精度の方が 10,000 件以降において高いもしくは変わらない結果を示した．これらより UNSW-NB15 において，一定以上まで教師データを無作為に選択することで，教師データ数が少ない状態における無作為選択による分類精度の方が高い問題の解決ができることが示された．

9. まとめ

侵入検知システムにおいて，教師データが少ない状態では無作為にデータ選択を行い，教師データ数が一定以上でデータ選択方法を能動学習に切り替えてデータ選択を行う手法を用いた分類モデルを評価した．評価では本手法のラベリングコスト削減への有効性を，侵入検知用ベンチマークデータセットである NSL-KDD と UNSW-NB15 を用いて確認した．さらに，能動学習を用いて機械学習型侵入検知システムを構築する際の問題点である，教師データ数が少ない状態において能動学習による分類精度が低いという問題を，データ選択方法を切り替える手法で解決可能であることの検証をも行った．検証により，データ選択方法を切り替える手法でも，データ選択方法を切り替えない能動学習と同様にラベリングコストの削減が可能であることを示した．また，教師データが少ない状態における分類精度の改善が可能であることが示唆された．

本研究では，データ選択方法を切り替える教師データ数を固定値にしたが，学習データによって適切な切り替え件数は異なることが分かった．そのため，今後の課題としてデータ選択方法切り替えを行う件数を自動的に決める手法の検討が挙げられる．

参考文献

[1] 赤坂省伍, 梅澤猛, 大澤範高, “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,” 情報処理学会第 79 回全国大会論文集, No.3, pp.591-592, 2017.

[2] Y. Bhavsar. and K. Waghmare., “ Intrusion Detection System Using Data Mining Technique: Support Vector Machine.,” International Journal of Emerging Technology and Advanced Engineering Volume 3, Issue 3, pp.581-586, 2013.

[3] S. Revathi. and A. Malathi., “A detailed analysis on NSL-KDD dataset using various machine learning techniques for intrusion detection.,” International Journal of Engineering Research and Technology. ESRSA Publications, 2013.

[4] B. Settles., “Active learning literature survey,” Computer Sciences Technical Report 1648, University of Wisconsin, 2010.

[5] University of New Brunswick, “NSL-KDD dataset,” ＜ http://www.unb.ca/cic/datasets/nsl.html ＞ , (2018/07/10 最終閲覧).

[6] N. Moustafa. and J. Slay., “The UNSW-NB15 data set description, ” <https://www.unsw.adfa.edu.au/unsw-canberra-cyber/cybersecurity/ADFA-NB15-Datasets/>, (2018/07/10 最終閲覧).

[7] N. Moustafa. and J. Slay., “UNSW-NB15: a comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set),” 2015 Military Communications and Information Systems Conference (MilCIS), IEEE, 2015.

[8] N. Moustafa. and J. Slay., “The evaluation of Network Anomaly Detection Systems: Statistical analysis of the UNSW-NB15 data set and the comparison with the KDD99 data set, ” Information Security Journal: A Global Perspective(2016): 1-14, 2016.

© 2018 Information Processing Society of Japan