6ZE-06 帰納論理プログラミングを用いた監査データからの知識抽出*

吉井 崇行† 小林 文彦† 溝口文雄†

東京理科大学 理工学部 経営工学科‡

1 はじめに

機器，または，ネットワークへの不正なアクセスは年々増加しており，侵入検知への期待が高まっている．その侵入検知には，統計や学習などを利用して正常な運用状態を把握し，例外を検出する異常検知 (anomaly detection) という研究領域がある．しかし，既存の異常検知システムでは，最適ルールの自動生成，自動更新，さらに，適用方法に関して検討段階である．そこで，本稿では，例からの学習が可能な帰納論理プログラミングシステム (ILP システム) をルール生成エンジンとし，これにパラメータチューニング機能を持たせることで，最適な性能を持つルールの自動生成を行なう．そして，それを利用した異常検知方法を検討する．

また，本研究は，カッコーエッグプロジェクト [1] に関連した研究である．

2 設計方針

本稿では，監査ログからルール生成するためのエンジンとして ILP システムを用いる．これは，Muggleton の逆伴意 (Inverse Entailment) [2] と溝口らによって開発された GKS [3] を参考にし，Java 言語により実装した学習システムである．この特徴は，正事例の包含率と仮説の長さを目的関数とし，負事例を含んでもよい割合 (エラー率) を制約条件として与え，ボトム節の中から最良仮説 (ルール) を見つける．

しかし，この ILP システムを含め，既存の ILP システムは，探索用パラメータを設定する必要がある．このため，従来の学習システムの利用者は，ルールの性能が満足いくまでこのパラメータを調整しなければならなかった．そこで，本稿では性能尺度の値 (パラメータ) に注目することで探索用パラメータの自動調整を行なう (パラメータチューニング) 機能を持たせる．さらに，この機能は，リサンプリング手法であるクロスバリデーションとブートストラップの結果を利用し，その学習結果からパラメータの調整を行なう．このように，リサンプリング手法を取り入れたことで，大規模ログデータに対しても効果的なルール生成が可能である．

そして，本稿では，このようにして作成されたルールを，図 1で示すようにプロファイリングフィルタとして利用し，生データとの適合性を見ることで異常検知を行なう．

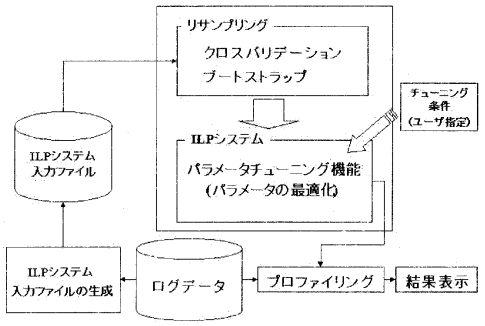

図 1: 検知の流れ

3 パラメータの自動調整方法

本稿で利用する ILP システムは，負事例をどれだけ含んでもよいかを示すエラー率を設定しなければならない．その設定の仕方によって，ルールの性能は左右される．そこで，ユーザが指定した性能尺度を最大にするように，エラー率を自動調整する．この時，指定されなかった残りのパラメータを制約条件として与えることができる．ここで，最大化できるパラメータは，Accuracy(精度)，Sensitivity(感度)，Specificity(特殊性) である．これらのパラメータは，リサンプリング手法から得られる [4].

また，感度はエラー率を大きくすると値が大きくなり，特殊性はエラーを小さくすると値が大きくなる傾向がある．そこで，パラメータの自動調整処理に関して，エラー率とパラメータの関係を考慮に入れた 2 分探索法により最良のパラメータ値を示す時のエラー率を求める．

4 実験

4.1 対象データ

研究室内の UNIX が動作している計算機 (ホスト)10 台に対して last コマンドを実行し，1999 年 1 月から 10 月までのログインログ 3,562 件を収集した．このログイ

*Knowledge mining from audit data using Inductive Logic Programming

†Takayuki Yoshii,Fumihiko Kobayashi,Fumio MIZOGUCHI

‡Department of Industrial Administration, Faculty of Sci. and Tech., Science University of Tokyo

ンログには,「あるユーザが，ある計算機に，いつ，どのようにログインしたか」という情報が記述されている．

4.2 学習方法

あるユーザを正常な利用者 (正事例) とし，それ以外のユーザを異常な利用者 (負事例) と仮定する．そして，その時のターミナル情報，ログイン元ホスト，ログイン先ホスト，曜日を背景知識に与え，各月ごとにルール生成を行なった．この時，クロスバリデーションにより事例を 10 分割した結果を利用し，チューニングにより最大化するパラメータを精度とし，その時の制約条件に感度を与えた．このように精度と感度に注目した条件を与えたのは，2 つの理由がある．1 つは，得られたルールがそのユーザ自身を正しく説明することができ，さらに，他のユーザの振る舞いをより良く区別するためである．もう 1 つは，精度と感度は，エラー率が高くなると，精度は下がり，感度は上がる傾向があり，例え，高い精度のルールが得られたからといって，感度が低ければ，他とはよく区別できるが，それ自体をうまく説明できないのは問題であると考えたからである．

4.3 学習結果

エラーを 7.5%認めた場合，精度は約 90%を示し，感度は約 89%を示した．この時のデフォルトの精度は，約 77%であったため，精度は非常に良い値を示している．つまり，得られたルールは，ユーザを正しくとらえることができ，さらに，他のユーザと区別することが可能であると言える．この時，以下のようなルールが得られた．

- ユーザ A は，ホスト X に対して ftp ログインする．
- ユーザ B は，ホスト X からホスト Y にログインする．

4.4 プロファイリングとしての適用

図 2と図 3は，ある月のルールを翌月のプロファイリングとして適用し，その時の適合度を求めた結果である．この適合度は以下のようにして求めた．

$$\text{適合度} = \frac{\text{ルールが真と判断した数}}{\text{テストしたデータの総数}}$$

ここで取り上げたユーザ N とユーザ H は，毎月，最低 20 回はログインし，他のユーザに比べて Unix の利用

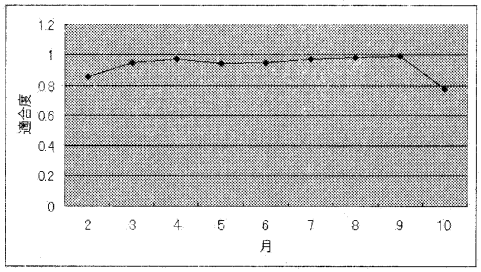

図 2: ユーザ N の場合

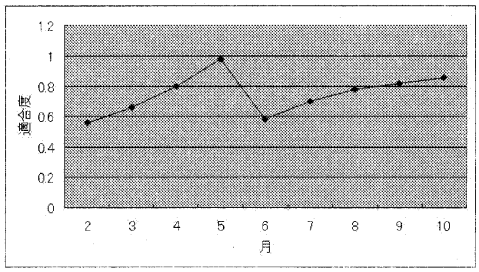

図 3: ユーザ H の場合

頻度の高いユーザである．しかし，ユーザ N は常に高い適合度を示した (図 2) が，ユーザ H は変化の激しい適合度を示した (図 3)．これらは，前者は毎月似通った振る舞いを行ない，後者は月によって振る舞いが変わることを意味している．

5 考察

本稿では，単一の監査ログに対して ILP システムを利用し，ルール生成を行なう方法を検討した．しかし，単一の監査ログのルールから異常検知を行なった場合，適合度が変動するユーザに対して，誤った判断をしてしまう可能性がある．そこで,「複数の監査ログによる複合検知」と「逐次による知識処理」について検討する必要があると思われる．前者は，あるログから得られたプロファイルが異常を示した場合，他のログ，または，他のルールを参照し，異常を確認することで，誤判断を減らすことを意味する．また，後者は，過去の事実の蓄積を考慮に入れることで，ユーザの行動の突然の変化に対して柔軟に対処できると思われる．この時，逐次学習可能な ILP システムの利用は有効であると考えられる．

6 まとめ

本稿では，監査ログに対して，パラメータの自動調整機能を備えた ILP システムを適用した．これにより，従来意識していた探索用パラメータの値を自動に求めることを可能にした．そして，このような機能を持つ ILP システムをルール生成エンジンとして利用し，ログインログを解析したところ，月によって振る舞いが変わるユーザと毎月一定したタスクを実行している 2 種類のユーザがいることが結果から得られた．

参考文献

[1] F.Mizoguchi,'Visual browser for intrusion detection - Cuckoo egg project', CSS'99, 1999.

[2] S.Muggleton,'Inverse Entailment and Progol', *New Generation Computing*,13:245-286, 1995.

[3] F.Mizoguchi and H.Ohwada,'Using Inductive Logic Programming for Constraint Acquistion in Constraint-based Problem Solving', Proc. of the 5 th International Workshop on ILP, 297-322, 1995.

[4] M.Weiss and A.Kulikowski, Computer Systems That Learn', Morgan Kaufmann Publishers, 1991.